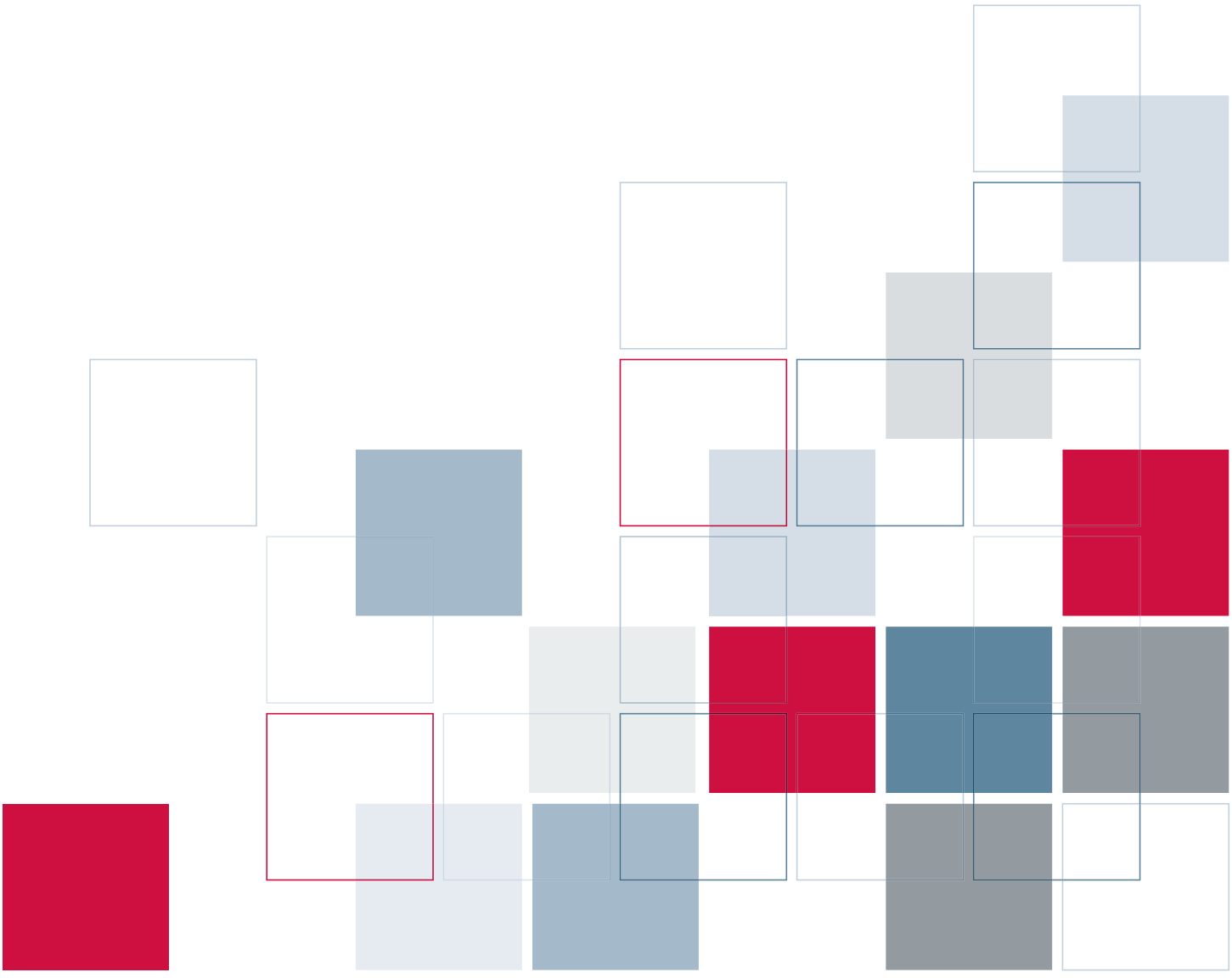
PASW® Missing Values (Fehlende Werte) 18



Weitere Informationen zu SPSS Inc.-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

SPSS ist eine eingetragene Marke.

PASW ist eine eingetragene Marke von SPSS Inc..

Die SOFTWARE und die Dokumentation werden mit BESCHRÄNKTEN RECHTEN zur Verfügung gestellt. Verwendung, Vervielfältigung und Veröffentlichung durch die Regierung unterliegen den Beschränkungen in Unterabschnitt (c)(1)(ii) von The Rights in Technical Data and Computer Software unter 52.227-7013. Vertragspartner/Hersteller ist SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.
Patentnr. 7.023.453

Allgemeiner Hinweis: Andere in diesem Dokument verwendete Produktnamen werden nur zu Identifikationszwecken genannt und können Marken der entsprechenden Unternehmen sein.

Windows ist eine eingetragene Marke der Microsoft Corporation.

Apple, Mac und das Mac-Logo sind Marken von Apple Computer, Inc., die in den USA und in anderen Ländern eingetragen sind.

Dieses Produkt verwendet WinWrap Basic, Copyright 1993–2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Kein Teil dieser Publikation darf ohne vorherige Einwilligung des Herstellers in irgendeiner Form elektronisch, mechanisch, durch Fotokopieren, Aufzeichnen oder sonstige Mittel reproduziert, in einem Datenabfragesystem gespeichert oder übertragen werden.

Vorwort

PASW Statistics 18 ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Fehlende Werte definieren bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Fehlende Werte definieren müssen zusammen mit PASW Statistics 18 Core verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von Fehlende Werte definieren Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS Inc. erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von Fehlende Werte definieren Erweiterungsmodul.

Kompatibilität

PASW Statistics kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS Inc.. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS Inc. wenden, um Informationen zum Kundendienst, zu Zahlungen oder zu Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Core-System ausgeliefert.

Kundendienst

Wenden Sie sich bei Fragen zur Lieferung oder Ihrem Kundenkonto an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS Inc. bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Seminaren wünschen, wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit PASW Statistics oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen zum Technischen Support finden Sie auf der Website unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Das Handbuch *SPSS Statistics Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für PASW Statistics 18 ist geplant. Das Handbuch *SPSS Statistics Advanced Statistical Procedures Companion*, bei dem auch PASW Statistics 18 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Statistics Guide to Data Analysis* für PASW Statistics 18 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Inhalt

Teil I: Benutzerhandbuch

1 Missing Values (Fehlende Werte) - Einleitung 1

2 Analyse fehlender Werte 3

Anzeige der Muster fehlender Werte 6
Anzeigen deskriptiver Statistiken für fehlende Werte 8
Schätzen von Statistiken und Imputieren fehlender Werte 9
EM-Schätzung: Optionen 10
Optionen für die Regressionsschätzung 12
Vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen) 13
Zusätzliche Funktionen beim Befehl MVA 14

3 Multiple Imputation 15

Muster analysieren. 17
Fehlende Datenwerte ersetzen 18
Methode 20
Nebenbedingungen 22
Ausgabe 24
Zusätzliche Funktionen beim Befehl MULTIPLE IMPUTATION 25
Arbeiten mit Daten aus multipler Imputation 25
Analysieren von Daten multipler Imputation 29
Multiple-Imputation-Optionen 34

Teil II: Beispiele

4 Analyse fehlender Werte 37

Beschreiben des Musters fehlender Daten 37

Durchführen der Analyse zur Anzeige deskriptiver Statistiken. 37
Evaluieren der deskriptiven Statistiken 39
Erneute Durchführung der Analyse zur Anzeige von Mustern. 45
Evaluieren der Mustertabelle 47
Erneute Durchführung der Analyse für den MCAR-Test nach Little 48

5 Multiple Imputation *50*

Verwendung von multipler Imputation für die Vervollständigung und Analyse einer Daten-Sets. . . 50
Analyse der Muster fehlender Werte 50
Automatische Imputation fehlender Werte. 54
Angepasstes Imputationsmodell 61
Prüfen auf FCS-Konvergenz 69
Analyse vollständiger Daten 73
Auswertung. 84

Anhang

A Beispieldateien *85*

Index *97*

Teil I: Benutzerhandbuch

Kapitel

1

Missing Values (Fehlende Werte) - Einleitung

Fälle mit fehlenden Werten stellen eine Herausforderung dar, da typische Modellverfahren diese Fälle einfach von der Analyse ausschließen. Wenn es wenige fehlende Werte (grob geschätzt weniger als 5 % der Gesamtzahl an Fällen) gibt und diese Werte als zufällig fehlend betrachtet werden, also das Fehlen eines Werts nicht von anderen Werten abhängt, dann ist die typische Methode des listenweisen Löschens relativ sicher. Die Option “Missing Values+” kann Ihnen helfen zu bestimmen, ob das listenweise Löschen ausreichend ist, und bietet anderenfalls Methoden zur Handhabung fehlender Werte.

Die Analyse fehlender Werte im Vergleich zu Verfahren multipler Imputation

Die Option “Missing Values” bietet zwei Arten von Verfahren für die Handhabung fehlender Werte:

- Die Verfahren der Multiplen Imputation bieten die Analyse von Mustern fehlender Daten und zielen auf eine eventuelle multiple Imputation der fehlenden Werte ab. Es werden mehrere Versionen des Daten-Sets erzeugt, von denen jede ein eigenes Set an imputierten Werten enthält. Wenn statistische Analysen durchgeführt werden, werden die Parameterschätzungen für alle imputierten Daten-Sets gesammelt. Sie bieten Schätzungen, die im Allgemeinen genauer als die einzelner Imputationen sind.
- Die Analyse fehlender Werte bietet ein geringfügig anderes Set an beschreibenden Tools für die Analyse fehlender Daten (im Besonderen den MCAR-Test von Little) und umfasst eine Vielzahl einfacher Imputationsmethoden. Beachten Sie, dass die multiple Imputation im Allgemeinen als der einzelnen Imputation überlegen betrachtet wird.

Aufgaben fehlender Werte

Sie können mit der Analyse fehlender Wert anhand der folgenden grundlegenden Schritte beginnen:

- **Untersuchen Sie das Fehlen.** Verwenden Sie die Analyse fehlender Werte und die Analyse von Mustern, um die Muster der fehlenden Werte in Ihren Daten zu untersuchen und zu bestimmen, ob eine multiple Imputation erforderlich ist.
- **Fehlende Werte vorschreiben.** Verwenden Sie “Fehlende Datenwerte ersetzen”, um imputierte fehlende Werte zu multiplizieren.
- **Analysieren Sie die “vollständigen Daten”.** Verwenden Sie ein Verfahren, das Daten der multiplen Imputation unterstützt. Informationen zur Analyse von Datensets der multiplen Imputation und

eine Liste der Verfahren, die diese Daten unterstützen, finden Sie unter Analysieren von Daten multipler Imputation auf S. 29.

Kapitel
2

Analyse fehlender Werte

Die Prozedur “Analyse fehlender Werte” dient primär drei Funktionen:

- Beschreiben des Musters fehlender Daten. Wo befinden sich die fehlenden Daten? Welches Ausmaß weisen sie auf? Tendieren Variablenpaare dazu, fehlenden Werte in mehreren Fällen aufzuweisen? Sind die Datenwerte extrem? Fehlen wahllos Werte?
- Schätzen der Mittelwerte, Standardabweichung, Kovarianzen und Korrelationen für verschiedene Methoden für fehlende Werte: listenweise, paarweise, Regression oder EM (Maximierung des Erwartungswerts). Bei der paarweisen Methode werden auch die Häufigkeiten der paarweise vollständigen Fälle angezeigt.
- Füllt (imputierte) fehlende Werte mit geschätzten Werten mithilfe von Regressions- oder EM-Methoden. Multiple Imputation wird in der Regel jedoch als Methode betrachtet, die die genaueren Ergebnisse liefert.

Die Analyse fehlender Werte unterstützt Sie beim Umgang mit Problemen, die durch unvollständige Daten verursacht werden. Wenn Fälle mit fehlenden Werten sich systematisch von Fällen ohne fehlende Werte unterscheiden, können die Ergebnisse irreführend sein. Fehlende Daten können außerdem die Genauigkeit der berechneten Statistiken beeinträchtigen, da weniger Informationen vorliegen als ursprünglich geplant. Ein weiteres Problem ist die Annahme hinter vielen statistischen Prozeduren, dass alle Fälle vollständig sind. Fehlende Werte können den erforderlichen theoretischen Ansatz verkomplizieren.

Beispiel. Bei der Auswertung einer Leukämiebehandlung werden verschiedene Variablen gemessen. Es sind jedoch nicht alle Messwerte für alle Patienten verfügbar. Die Muster der fehlenden Daten werden angezeigt, tabellarisch dargestellt und für zufällig befunden. Eine EM-Analyse wird für die Schätzung der Mittelwerte, Korrelationen und Kovarianzen verwendet. Sie dient außerdem dazu, um festzustellen, ob die Daten in völlig zufälliger Weise fehlen. Die fehlenden Werte werden dann durch abgeleitete (imputierte) Werte ersetzt und zur weiteren Analyse in einer neuen Datendatei gespeichert.

Statistiken. Univariate Statistiken, einschließlich der Anzahl nichtfehlender Werte, dem Mittelwert, der Standardabweichung, der Anzahl fehlender Werte und der Anzahl von Extremwerten. Geschätzte Mittelwerte, Kovarianz- und Korrelationsmatrix unter Verwendung der listenweisen, paarweisen, EM- oder Regressionsmethode. MCAR-Test nach Little mit EM-Ergebnissen. Auswertung der Mittelwerte nach verschiedenen Methoden. Für Gruppen, die durch fehlende gegenüber nichtfehlende Werte definiert sind: *T*-Tests. Für alle Variablen: Muster der fehlenden Werte angezeigt nach Fällen und Variablen.

Erläuterung der Daten

Daten. Die Daten können kategorial oder quantitativ (metrisch oder stetig) sein. Die Berechnung von Statistiken und das Vorschreiben (Imputieren) fehlender Daten ist jedoch nur für die quantitativen Variablen möglich. Bei allen Variablen müssen die fehlenden Werte, die nicht als systemdefiniert fehlend kodiert sind, als benutzerdefiniert fehlend definiert werden. Wenn beispielsweise für eine Frage in einem Fragebogen die Antwort *Ich weiß nicht* als 5 kodiert ist und Sie diese als fehlend behandeln möchten, muss für diese Frage 5 als benutzerdefinierter fehlender Wert kodiert werden.

Annahmen. Listenweisen, paarweisen und Regressionsschätzungen liegt die Annahme zugrunde, dass das Muster der fehlenden Werte nicht von den Datenwerten abhängt. Diese Bedingung ist als **völlig zufällig fehlend** oder MCAR ("missing completely at random") bekannt. Daher ergeben alle Schätzmethoden (einschließlich der EM-Methode) bei MCAR-Daten konsistente und unverzerrte Schätzer der Korrelationen und Kovarianzen. Die Verletzung der MCAR-Annahme kann dazu führen, dass von der listenweisen, paarweisen bzw. Regressionsmethode verzerrte Schätzer generiert werden. Wenn es sich nicht um MCAR-Daten handelt, muss die EM-Schätzung verwendet werden.

Der EM-Schätzung liegt die Annahme zugrunde, dass das Muster der fehlenden Daten nur mit den beobachteten Daten zusammenhängt. Diese Bedingung wird **zufällig fehlend** oder MCAR ("missing at random") genannt. Aufgrund dieser Annahme können die Schätzungen unter Verwendung der verfügbaren Information korrigiert werden. So kann es beispielsweise in einer Studie über Bildung und Einkommen vorkommen, dass bei Personen mit niedrigerer Bildung eine höhere Anzahl fehlende Einkommenswerte vorliegt. In diesem Fall handelt es sich um MAR-Daten, nicht um MCAR-Daten. Anders ausgedrückt: Bei MAR hängt die Wahrscheinlichkeit, dass ein Einkommen angegeben wird, vom Bildungsniveau der betreffenden Person ab. Die Wahrscheinlichkeit kann abhängig von der Bildung, nicht jedoch abhängig vom Einkommen *innerhalb des betreffenden Bildungsniveaus* schwanken. Wenn die Wahrscheinlichkeit, dass ein Einkommen angegeben wird auch in Abhängigkeit vom Einkommen innerhalb der einzelnen Bildungsniveaus schwankt (wenn beispielsweise Personen mit hohem Einkommen ihr Einkommen nicht angeben), handelt es sich weder um MCAR-Daten noch um MAR-Daten. Dies ist eine ungewöhnliche Situation, bei deren Eintreten keine der Methoden angemessen ist.

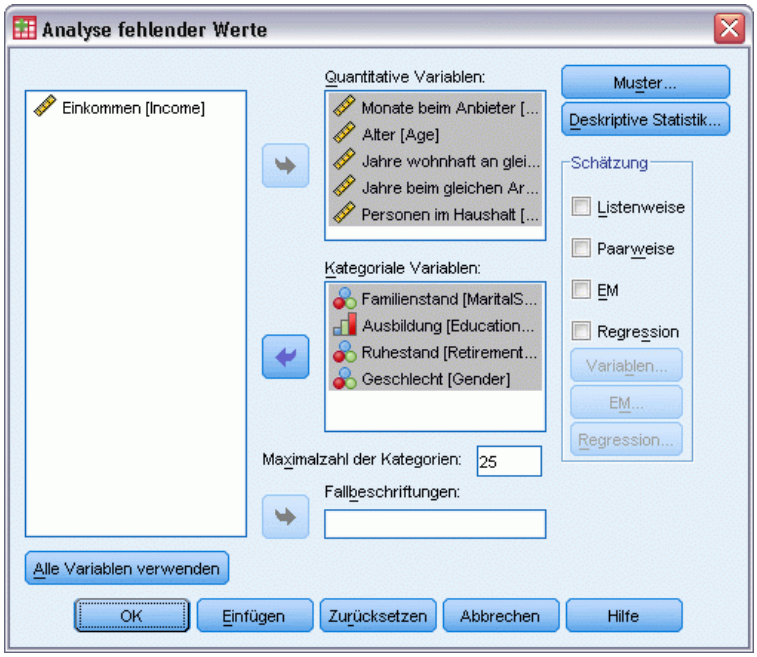
Verwandte Prozeduren. Listenweise und paarweise Schätzungen können in vielen Prozeduren verwendet werden. Mit der linearen Regression und der Faktorenanalyse könne fehlende Werte durch die Mittelwerte ersetzt werden. Im Erweiterungsmodul "Forecasting" sind verschiedene Methoden verfügbar, um fehlende Werte in Zeitreihen zu ersetzen.

So berechnen Sie eine Analyse fehlender Werte:

► Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Analyse fehlender Werte...

Abbildung 2-1
Dialogfeld "Analyse fehlender Werte"



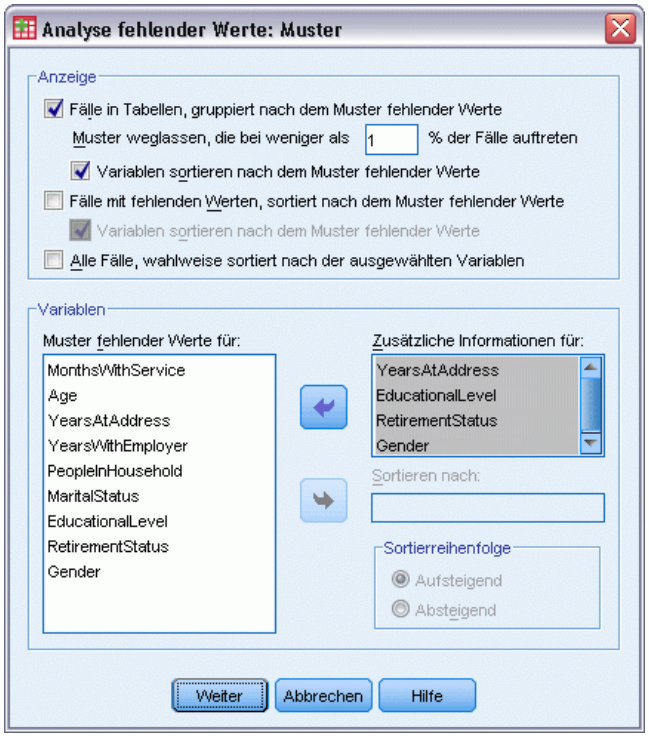
- ▶ Wählen Sie mindestens eine quantitative (metrische) Variable zur Schätzung der Statistiken und der optionalen Imputation fehlender Werte aus.

  Die folgenden Optionen sind verfügbar:

  - Wählen Sie kategoriale Variablen (numerisch oder String) aus und geben Sie für die Anzahl der Kategorien eine Grenze (Maximalzahl der Kategorien) ein.
  - Klicken Sie auf Muster zur tabellarischen Darstellung der Muster fehlender Daten. Für weitere Informationen siehe Anzeige der Muster fehlender Werte auf S. 6.
  - Klicken Sie auf Deskriptive Statistik zur Anzeige deskriptiver Statistiken fehlender Werte. Für weitere Informationen siehe Anzeigen deskriptiver Statistiken für fehlende Werte auf S. 8.
  - Wählen Sie eine Methode zur Schätzung der Statistiken (Mittelwerte, Kovarianzen und Korrelationen) und optionalen Imputation fehlender Werte aus. Für weitere Informationen siehe Schätzen von Statistiken und Imputieren fehlender Werte auf S. 9.
  - Wenn Sie "EM" oder "Regression" auswählen, klicken Sie auf Variablen, um die Untergruppe anzugeben, die für die Schätzung verwendet wird. Für weitere Informationen siehe Vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen) auf S. 13.
  - Wählen Sie eine Variable für die Fallbeschriftung aus. Diese Variable dient zur Beschriftung von Fällen in Mustertabellen, die einzelne Fälle anzeigen.

Anzeige der Muster fehlender Werte

Abbildung 2-2
Dialogfeld "Analyse fehlender Werte: Muster"



Sie können verschiedene Tabellen anzeigen lassen, die die Muster und das Ausmaß der fehlenden Daten zeigen. Mit diesen Tabellen können Sie Antworten auf folgende Fragen finden:

- Wo befinden sich fehlende Daten?
- Tendieren Variablenpaare dazu, fehlende Werte in einzelnen Fällen aufzuweisen?
- Sind Datenwerte extrem?

Anzeigen

Für die Anzeige von Mustern fehlender Daten stehen drei Tabellentypen zur Verfügung.

Fälle in Tabellen. Die Muster fehlender Daten in den Analysevariablen werden in Tabellenform dargestellt, wobei für jedes Muster auch die Häufigkeiten angegeben werden. Mit Variable sortieren nach dem Muster fehlender Werte können Sie angeben, ob Häufigkeiten (Anzahl) und Variablen nach der Ähnlichkeit der Muster sortiert werden sollen. Mit Muster weglassen bei weniger als n % der Fälle können Sie Muster ausschließen, die nur selten vorkommen..

Fälle mit fehlenden Werten. Für die einzelnen Analysevariablen werden jeweils die einzelnen Fälle mit einem fehlenden Wert oder einem Extremwert tabellarisch dargestellt. Mit Variable sortieren nach dem Muster fehlender Werte können Sie angeben, ob Häufigkeiten (Anzahl) und Variablen nach der Ähnlichkeit der Muster sortiert werden sollen.

Alle Fälle. Die einzelnen Fälle werden tabellarisch dargestellt, und fehlende Werte und Extremwerte werden für jede Variable angegeben. Die Fälle werden in der Reihenfolge aufgeführt, in der sie in der Datendatei auftreten, sofern unter Sortieren nach keine Variable angegeben wurde.

In den Tabellen, die einzelne Fälle anzeigen, werden folgende Symbole verwendet:

+	Extrem hoher Wert
-	Extrem niedriger Wert
F	Systemdefiniert fehlender Wert
D	Erster Typ des benutzderdefinierten fehlenden Werts
K	Zweiter Typ des benutzderdefinierten fehlenden Werts
R	Dritter Typ des benutzerdefinierten fehlenden Werts

Variablen

Sie können weitere Informationen für die in die Analyse aufgenommenen Variablen anzeigen. Die Variablen, die Sie unter Zusätzliche Informationen für hinzufügen, werden einzeln in der Tabelle der fehlenden Muster angezeigt. Bei quantitativen (metrischen) Variablen wird der Mittelwert und bei kategorialen Variablen wird die Anzahl der Fälle aufgeführt, die das Muster in jeder Kategorie aufweisen.

- **Sortieren nach.** Die Fälle werden entsprechend der aufsteigenden oder absteigenden Reihenfolge der Werte der angegebenen Variablen aufgeführt. Diese Option ist nur für Alle Fälle verfügbar.

So geben Sie Muster fehlender Werte an:

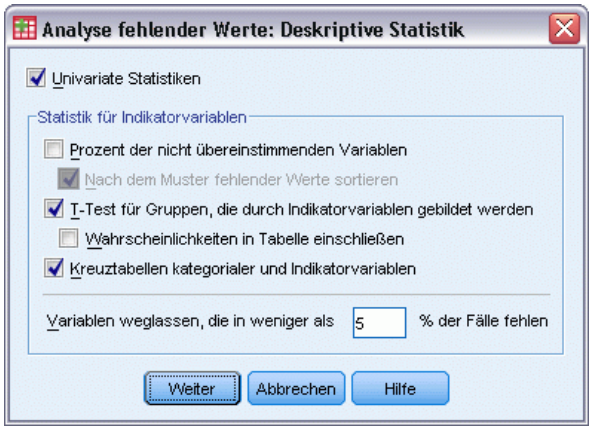
▶ Wählen Sie im Dialogfeld “Analyse fehlender Werte” die Variablen aus, für die Muster fehlender Werte angezeigt werden sollen.

▶ Klicken Sie auf Muster.

▶ Wählen Sie die anzuzeigenden Mustertabellen aus.

Anzeigen deskriptiver Statistiken für fehlende Werte

Abbildung 2-3
Dialogfeld "Analyse fehlender Werte: Deskriptive Statistik"



Univariate Statistiken

Univariate Statistiken können zur Ermittlung des allgemeinen Ausmaßes der fehlenden Daten beitragen. Für jede Variable werden folgende Daten angezeigt:

- Anzahl nichtfehlender Werte
- Anzahl und Prozentsatz fehlender Werte.

Für quantitative (metrische) Variablen werden außerdem folgende Daten angezeigt:

- Mittelwert
- Standardabweichung
- Anzahl extrem hoher und niedriger Werte

Statistik für Indikatorvariablen

Für jede Variable wird eine Indikatorvariable erstellt. Diese kategoriale Variable gibt an, ob die Variable für einen einzelnen Fall vorhanden ist oder fehlt. Die Indikatorvariablen werden verwendet, um die Tabellen mit Nichtübereinstimmungen, *T*-Tests und Häufigkeiten zu erstellen.

Prozent der nicht übereinstimmenden Variablen. Für jedes Variablenpaar wird der Prozentsatz von Fällen angezeigt, in denen eine Variable einen fehlenden Wert und die andere Variable einen nichtfehlenden Wert aufweist. Jedes diagonale Element in der Tabelle enthält den Prozentsatz von fehlenden Werten für eine einzelne Variable.

T-Test für Gruppen, die durch Indikatorvariablen gebildet werden. Für jede quantitative Variable werden die Mittelwerte von zwei Gruppen mithilfe der Student-*T*-Statistik verglichen. Die Gruppen geben an, ob eine Variable vorhanden ist oder fehlt. Es werden die *T*-Statistik, Freiheitsgrade, Häufigkeiten von fehlenden und nichtfehlenden Werten sowie die Mittelwerte der beiden Gruppen angezeigt. Außerdem können Sie alle zweiseitigen Wahrscheinlichkeiten anzeigen, die der *T*-Statistik zugeordnet sind. Wenn Ihre Analyse zu mehreren Tests führt, dürfen

Sie diese Wahrscheinlichkeiten nicht für Signifikanztests verwenden. Die Wahrscheinlichkeiten sind nur geeignet, wenn nur ein einziger Test berechnet wird.

Kreuztabellen kategorialer und Indikatorvariablen. Für jede kategoriale Variable wird eine Tabelle angezeigt. In der Tabelle werden für jede Kategorie die Häufigkeit und der Prozentsatz von nichtfehlenden Werten für die anderen Variablen angezeigt. Außerdem werden die Prozentsätze für jeden Typ von fehlenden Werten angezeigt.

Variablen weglassen, die in weniger als n % der Fälle fehlen. Um die Tabellen zu verkleinern, können Sie die Statistiken weglassen, die nur für eine kleine Anzahl von Fällen berechnet werden.

So zeigen Sie deskriptive Statistiken an:

- Wählen Sie im Dialogfeld "Analyse fehlender Werte" die Variablen aus, für die deskriptive Statistiken fehlender Werte angezeigt werden sollen.
- Klicken Sie auf Deskriptive Statistik.
- Wählen Sie die anzuzeigende deskriptive Statistik aus.

Schätzen von Statistiken und Imputieren fehlender Werte

Sie können Mittelwerte, Standardabweichung, Kovarianzen und Korrelationen unter Verwendung der listenweisen Methode (nur vollständige Fälle), der paarweisen Methode, der EM-Methode (Maximierung des Erwartungswerts) bzw. der Regressionsmethode schätzen. Außerdem können Sie auswählen, dass die fehlenden Werte imputiert (vorgeschrieben) werden sollen, d. h. dass Ersatzwerte geschätzt werden sollen. Beachten Sie, dass Multiple Imputation im Allgemeinen bei der Lösung des Problems fehlender Werte der einfachen Imputation überlegen ist. Der MCAR-Test von Little ist nach wie vor hilfreich bei der Bestimmung, ob eine Imputation erforderlich ist.

Listenweise Methode

Bei dieser Methode werden nur vollständige Fälle verwendet. Wenn eine der Analysevariablen fehlende Werte aufweist, wird der betreffende Fall aus den Berechnungen ausgeschlossen.

Paarweise Methode

Bei dieser Methode werden Paare von Analysevariablen betrachtet und ein Fall wird nur verwendet, wenn er für beide Variablen nichtfehlende Werte aufweist. Häufigkeiten, Mittelwerte und Standardabweichungen werden für jedes Paar gesondert berechnet. Da andere fehlende Werte im Fall ignoriert werden, sind die für zwei Variablen berechneten Korrelationen und Kovarianzen nicht von Werten abhängig die in anderen Variablen fehlen.

EM-Methode

Bei dieser Methode wird von einer Verteilung für die teilweise fehlenden Daten ausgegangen und die Schlussfolgerungen (Inferenzen) beruhen auf der Likelihood bei dieser Verteilung. Jede Iteration besteht aus einem E-Schritt und einem M-Schritt. Im E-Schritt wird die bedingte Erwartung der "fehlenden" Daten ermittelt, die auf den beobachteten Werten und den aktuellen

Schätzern der Parameter beruht. Anschließend werden die “fehlenden” Daten durch diese Erwartungen ersetzt. Im M-Schritt werden Maximum-Likelihood-Schätzer der Parameter so berechnet, wie wenn die fehlenden Daten ergänzt worden wären. “Fehlend” steht in Anführungszeichen, da die fehlenden Werte nicht direkt ergänzt werden. Stattdessen, werden bei der Log-Likelihood Funktionen dieser Werte verwendet.

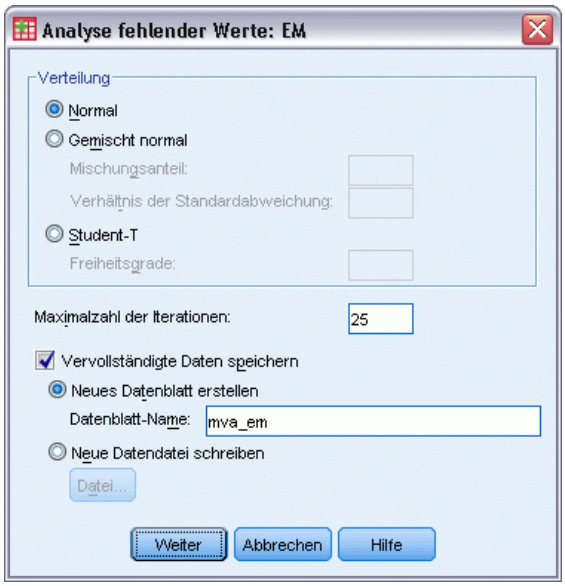
Die Chi-Quadrat-Statistik nach Roderick J. A. Little, die dazu dient zu testen, ob Werte in völlig zufälliger Weise fehlen (missing completely at random – MCAR) ist als Fußnote zu den EM-Matrizen abgedruckt. Bei diesem Test besagt die Nullhypothese, dass die Daten völlig zufällig fehlen, und der *p*-Wert ist auf dem Niveau 0,05 signifikant. Wenn der Wert weniger als 0,05 beträgt, fehlen die Werte nicht völlig zufällig. Die Daten fehlen möglicherweise zufällig (missing at random – MAR) oder fehlen nicht zufällig (missing at random – NMAR). Sie können nicht von einer der Eigenschaften ausgehen, sondern müssen die Daten analysieren, um zu ermitteln, in welcher Form sie fehlen.

Regressionsmethode (Factor Analysis)

Diese Methode berechnet Schätzer für die mehrfach lineare Regression und verfügt über Optionen zur Erweiterung der Schätzer durch Zufallskomponenten. Zu jedem vorhergesagten Wert kann das Verfahren ein Residuum aus einem zufällig ausgewählten vollständigen Fall, eine normale Zufallsabweichung oder eine Zufallsabweichung (anhand der Quadratwurzel der Residualvarianz (residual mean square) aus der *t*-Verteilung hinzufügen.

EM-Schätzung: Optionen

Abbildung 2-4
Dialogfeld “Analyse fehlender Werte: EM”



Beim EM-Verfahren werden unter Verwendung eines iterativen Prozesses die Mittelwerte, die Kovarianzmatrix und die Korrelation der quantitativen (metrischen) Variablen mit fehlenden Werte geschätzt.

Verteilung. EM erstellt Schlussfolgerungen (Inferenzen) anhand der für die jeweilige Verteilung geltenden Likelihood. Standardmäßig wird eine Normalverteilung angenommen. Wenn Sie wissen, dass die Flanken der Verteilung länger sind als die einer Normalverteilung, können Sie anfordern, dass die Prozedur die Likelihood-Funktion aus einer Student-*T*-Verteilung mit *n* Freiheitsgraden erstellt. Die gemischte Normalverteilung führt ebenfalls zu einer Verteilung mit längeren Flanken. Geben Sie die Quotienten der Standardabweichungen der gemischten Normalverteilung und das Mischungsverhältnis der beiden Verteilungen an. Bei der gemischten Normalverteilung wird davon ausgegangen, dass nur die Standardabweichungen der Verteilungen unterschiedlich sind. Die Mittelwerte müssen übereinstimmen.

Maximale Anzahl der Iterationen. Legt die maximale Anzahl der Iterationen zur Schätzung der wahren Kovarianz fest. Die Prozedur wird beendet, wenn diese Anzahl der Iterationen erreicht wurde, auch wenn die Schätzer nicht konvergiert haben.

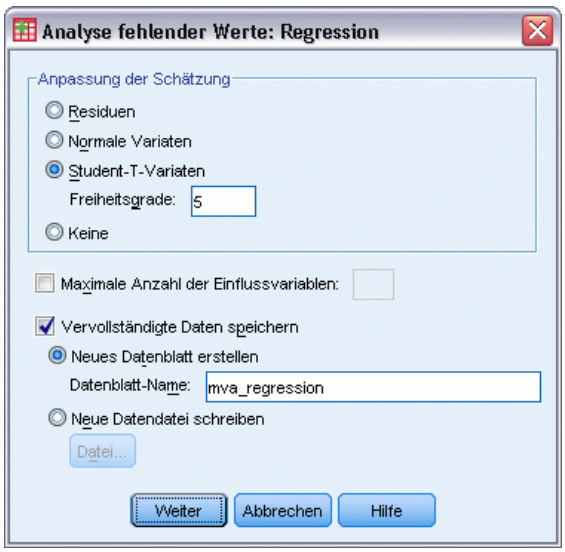
Vervollständigte Daten speichern. Sie können ein Daten-Set mit den imputierten Werten anstelle der fehlenden Werte speichern. Beachten Sie jedoch, dass kovarianzbasierte Statistiken, die die imputierten Werte verwenden, zu einer zu niedrige Schätzung der zugehörigen Parameterwerte führen. Der Grad der Unterschätzung ist proportional zu der Anzahl der Fälle die gemeinsam unbeobachtet sind.

So legen Sie EN-Optionen fest:

- Wählen Sie im Dialogfeld “Analyse fehlender Werte” die Variablen aus, für die fehlende Werte mithilfe der EM-Methode geschätzt werden sollen.
- Aktivieren Sie im Gruppenfeld “Schätzung” die Option EM.
- Klicken Sie auf die Schaltfläche Variablen, um die vorhergesagten Variablen und die Einflußvariablen anzugeben. Für weitere Informationen siehe Vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen) auf S. 13.
- Klicken Sie auf EM.
- Wählen Sie die gewünschten EM-Optionen aus.

Optionen für die Regressionsschätzung

Abbildung 2-5
Dialogfeld "Analyse fehlender Werte: Regression"



Bei der Regressionsmethode werden fehlende Werte unter Verwendung der mehrfachen linearen Regression geschätzt. Es werden die Mittelwerte, die Kovarianzmatrix und die Korrelationsmatrix der vorhergesagten Variablen angezeigt.

Anpassung der Schätzung. Bei der Regression kann den Regressionsschätzern eine Zufallskomponente hinzugefügt werden. Sie können Residuen, normale Variaten, Student-*T*-Variaten oder keine Anpassung auswählen.

- **Residuen.** Es werden Fehlerterme zufällig aus den beobachteten Residuen vollständiger Fälle ausgewählt und zu den Regressionsschätzungen addiert.
- **Normale Variaten.** Fehlerterme werden beliebig aus einer Verteilung mit dem Erwartungswert 0 und einer Standardabweichung gleich der Quadratwurzel der mittleren Quadratsumme des Regressionsfehlerterms gezogen.
- **Student-T-Variaten.** Fehlerterme werden beliebig aus der t(n)-Verteilung gezogen und anhand der Wurzel des mittleren Fehlerquadrats (RMSE) skaliert.

Maximale Anzahl der Einflussvariablen. Legt eine Obergrenze für die Anzahl der (unabhängigen) Einflußvariablen fest, die bei der Schätzung verwendet werden.

Vervollständigte Daten speichern. Schreibt ein Daten-Set in der aktuellen Sitzung oder eine externe Datendatei im PASW Statistics-Format. Dabei werden die fehlenden Werte durch die Werte ersetzt, die bei der Regression geschätzt wurden.

So legen Sie die Regressionsoptionen fest:

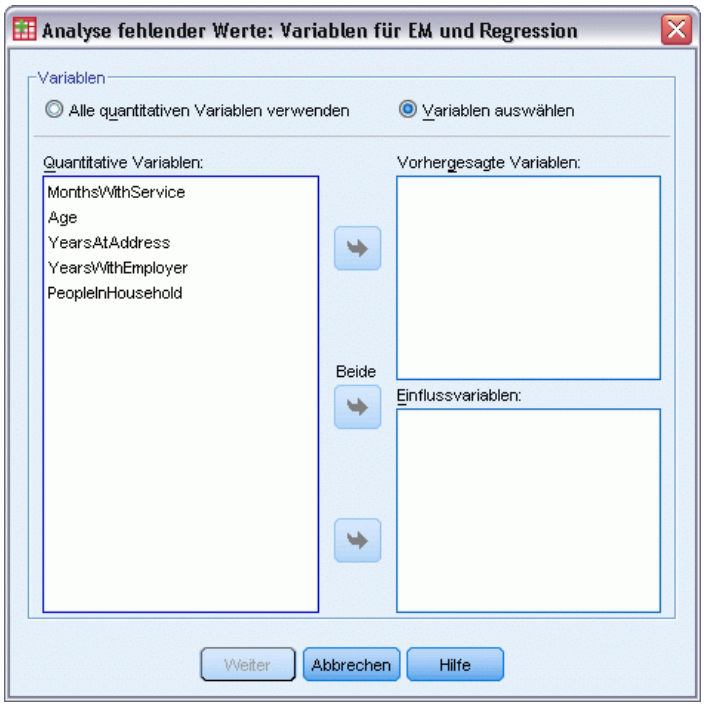
▶ Wählen Sie im Dialogfeld "Analyse fehlender Werte" die Variablen aus, für die fehlende Werte mithilfe der Regressionsmethode geschätzt werden sollen.

▶ Aktivieren Sie im Gruppenfeld "Schätzung" die Option Regression.

- Klicken Sie auf die Schaltfläche Variablen, um die vorhergesagten Variablen und die Einflußvariablen anzugeben. Für weitere Informationen siehe Vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen) auf S. 13.
- Klicken Sie auf Regression.
- Wählen Sie die gewünschten Regressionsoptionen aus.

Vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen)

Abbildung 2-6
Dialogfeld "Analyse fehlender Werte: Variablen für EM und Regression"



Standardmäßig werden alle quantitativen Variablen für EM und Regressionsschätzung verwendet. Falls erforderlich, können Sie bestimmte Variablen als vorhergesagte Variablen bzw. Einflussvariablen in den Schätzungen auswählen. Eine Variable kann prinzipiell in beiden Listen enthalten sein, es gibt jedoch Situationen, in denen es sinnvoll ist, die Verwendung einer Variablen einzuschränken. So vermeiden es einige Analytiker, die Werte von Ergebnisvariablen zu schätzen. Außerdem kann es sinnvoll sein, für verschiedene Schätzungen auch unterschiedliche Variablen zu verwenden und die Prozedur mehrmals auszuführen. Wenn Ihnen beispielsweise ein Set von Items vorliegt, bei denen es sich um die Bewertungen des Pflegepersonals handelt, und ein weiteres Set mit den Bewertungen der Ärzteschaft, kann es sinnvoll sein, eine Ausführung zur Schätzung der fehlenden Items für das Pflegepersonal und eine weitere Ausführung für die Schätzer der Items der Ärzteschaft durchzuführen.

Bei Verwendung der Regressionsmethode ist noch ein weiterer Faktor zu berücksichtigen. Bei der mehrfachen Regression kann die Verwendung einer großen Untergruppe unabhängiger Variablen zu schlechteren vorhergesagten Werten führen als eine kleinere Untergruppe. Daher

muss eine Variable mindestens ein *F* für die Aufnahme von 4,0 erreichen, um verwendet zu werden. Dieser Grenzwert kann über die Syntax geändert werden.

So geben Sie vorhergesagte Variablen und Vorhersagevariablen (Einflussvariablen) an:

- Wählen Sie im Dialogfeld "Analyse fehlender Werte" die Variablen aus, für die fehlende Werte mithilfe der Regressionsmethode geschätzt werden sollen.
- Aktivieren Sie im Gruppenfeld "Schätzung" die Option EM oder Regression.
- Klicken Sie auf Variablen.
- Wenn Sie nur bestimmte und nicht alle Variablen als vorhergesagte Variablen und Einflussvariablen verwenden möchten, aktivieren Sie Variablen auswählen und verschieben Sie die Variablen in die entsprechende(n) Liste(n).

Zusätzliche Funktionen beim Befehl MVA

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Schlüsselwort `DESCRIBE` in den Unterbefehlen `MPATTERN`, `DPATTERN` und `TPATTERN` können Sie separate deskriptive Variablen für Muster fehlender Werte, Datenmuster und Muster in Tabellen festlegen.
- Mit dem Unterbefehl `DPATTERN` können Sie mehrere Sortiervariablen für die Tabelle der Datenmuster festlegen.
- Mit dem Unterbefehl `DPATTERN` können Sie mehrere Sortiervariablen für die Datenmuster festlegen.
- Mit dem Unterbefehl `EM` können Sie die Toleranz und Konvergenz festlegen.
- Mit dem Unterbefehl `REGRESSION` können Sie die Toleranz und den *F*-Wert für die Aufnahme festlegen.
- Mit den Unterbefehlen `EM` und `REGRESSION` können Sie verschiedene Variablenlisten für das EM-Verfahren und die Regression festlegen.
- Für `TTESTS`, `TABULATE` und `MISMATCH` können Sie unterschiedliche Prozentsätze für das Unterdrücken von angezeigten Fällen festlegen.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
3

Multiple Imputation

Der Zweck der multiplen Imputation ist die Erzeugung möglicher Werte für fehlende Werte, um so verschiedene “vollständige” Sets an Daten zu erzeugen. Analyseverfahren, die mit Datensets aus multipler Imputation arbeiten, erzeugen Ausgaben für jedes “vollständige” Daten-Set sowie eine gemeinsame Ausgabe, die schätzt, welche Ergebnisse entstanden wären, wenn das Original-Daten-Set keine fehlenden Werte besitzen würde. Diese gemeinsamen Ergebnisse sind in der Regel genauer als die, die durch einfache Imputationsmethoden entstehen.

Analysevariablen. Die Analysevariablen können wie folgt gestaltet sein:

- **Nominal.** Eine Variable kann als nominal behandelt werden, wenn ihre Kategorien sich nicht in eine natürliche Reihenfolge bringen lassen, z. B. die Firmenabteilung, in der eine Person arbeitet. Beispiele für nominale Variablen sind Region, Postleitzahl oder Religionszugehörigkeit.
- **Ordinal.** Eine Variable kann als ordinal behandelt werden, wenn ihre Werte für Kategorien stehen, die eine natürliche Reihenfolge aufweisen (z. B. Grad der Zufriedenheit mit Kategorien von sehr unzufrieden bis sehr zufrieden). Ordinale Variablen treten beispielsweise bei Einstellungsmessungen (Zufriedenheit oder Vertrauen) und bei Präferenzbeurteilungen auf.
- **Metrisch.** Eine Variable kann als metrisch (stetig) behandelt werden, wenn ihre Werte geordnete Kategorien mit einer sinnvollen Metrik darstellen, sodass man sinnvolle Aussagen über die Abstände zwischen den Werten machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

  Bei der Prozedur wird davon ausgegangen, dass allen Variablen das richtige Messniveau zugewiesen wurde. Sie können das Messniveau für eine Variable jedoch vorübergehend ändern. Klicken Sie hierzu mit der rechten Maustaste auf die Variable in der Liste der Quellvariablen und wählen Sie das gewünschte Messniveau im Kontextmenü aus.

Messniveau und Datentyp sind durch ein Symbol neben der jeweiligen Variablen in der Variablenliste gekennzeichnet:

Messniveau	**Datentyp**			
	Numerisch	Zeichenfolge	Datum	Zeit
Metrisch (stetig)		entfällt		

Ordinal				
Nominal				

Häufigkeitsgewichtungen. Häufigkeitsgewichtungen (Replikation) werden von dieser Prozedur berücksichtigt. Fälle mit einer negativen oder nullwertigen Replikationsgewichtung werden ignoriert. Nichtganzzahlige Gewichtungen werden auf die nächste Ganzzahl gerundet.

Analysegewichtung. Analysegewichtungen (Regression oder Stichprobe) werden in Zusammenfassungen von fehlenden Werten und in passende Imputationsmodelle integriert. Fälle mit einer negativen oder nullwertigen Analysegewichtung werden ausgeschlossen.

Komplexe Stichproben. Das Verfahren der multiplen Imputation ist nicht explizit für Schichten, Cluster oder andere komplexe Stichprobenstrukturen gedacht, es kann jedoch endgültige Stichprobengewichtungen in Form der Analysegewichtungsvariablen akzeptieren. Beachten Sie auch, dass Prozeduren für komplexe Stichproben nicht automatisch mehrere imputierte Daten-Sets analysieren. Eine komplette Liste der Prozeduren, die Pooling unterstützen, finden Sie unter Analysieren von Daten multipler Imputation auf S. 29.

Fehlende Werte. Sowohl benutzer- als auch systemdefiniert fehlende Werte werden als ungültige Werte behandelt. Beide Arten von fehlenden Werten werden ersetzt, wenn Werte imputiert werden, und beide Arten werden als ungültige Werte von als Einflussfaktoren in Imputationsmodellen verwendeten Variablen behandelt. Benutzer- und systemdefiniert fehlende Werte werden auch bei Fehlanalysen als fehlende Werte behandelt.

Replikation von Ergebnissen (Fehlende Datenwerte ersetzen). Wenn Sie Ihre Imputation exakt reproduzieren möchten, müssen Sie nicht nur dieselben Einstellungen für die Prozedur, sondern auch denselben Initialisierungswert für den Zufallszahlengenerator, dieselbe Datenreihenfolge und dieselbe Variablenreihenfolge verwenden.

- **Generierung von Zufallszahlen.** Die Prozedur verwendet Zufallszahlengenerierung bei der Berechnung der imputierten Werte. Um zu einem späteren Zeitpunkt dieselben randomisierten Ergebnisse zu reproduzieren, müssen Sie vor jeder Ausführung der Prozedur “Fehlende Datenwerte ersetzen” denselben Initialisierungswert für den Zufallszahlengenerator verwenden.
- **Fallreihenfolge.** Werte werden in der Fallreihenfolge imputiert.
- **Reihenfolge der Variablen.** Die Imputationsmethode der vollständig konditionalen Spezifikation imputiert Werte in der Reihenfolge der Liste der Analysevariablen.

Für multiple Imputation stehen zwei spezielle Dialogfelder zur Verfügung.

- Muster analysieren bietet deskriptive Messungen der Muster von fehlenden Werten in den Daten und eignet sich als Untersuchungsschritt vor der Imputation.
- Fehlende Datenwerte ersetzen wird verwendet, um multiple Imputationen zu erzeugen. Die vollständigen Daten-Sets können mit Prozeduren analysiert werden, die Daten-Sets mit multipler Imputation unterstützen. Informationen zur Analyse von Datensets der multiplen Imputation und eine Liste der Verfahren, die diese Daten unterstützen, finden Sie unter Analysieren von Daten multipler Imputation auf S. 29.

Muster analysieren

"Muster analysieren" bietet deskriptive Messungen der Muster der fehlenden Werte in den Daten und eignet sich als Untersuchungsschritt vor der Imputation.

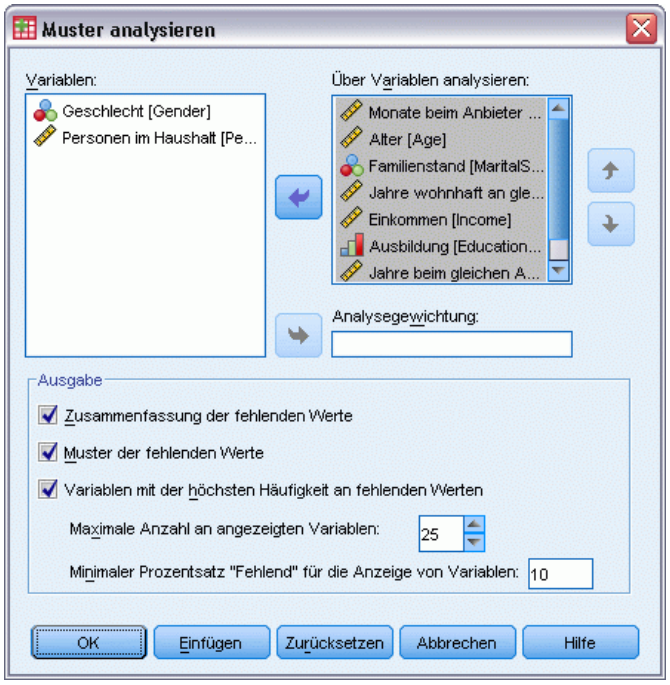
Beispiel. Ein Telekommunikationsanbieter möchte einen besseren Einblick in die Servicenutzungsmuster in seiner Kundendatenbank gewinnen. Er verfügt über die vollständigen Daten der von seinen Kunden genutzten Services, jedoch fehlen in den demographischen Informationen, die das Unternehmen gesammelt hat, einige Werte. Eine Analyse der Muster von fehlenden Werten kann helfen, die nächsten Schritte für die Imputation zu bestimmen. Für weitere Informationen siehe Verwendung von multipler Imputation für die Vervollständigung und Analyse einer Daten-Sets in Kapitel 5 auf S. 50.

So analysieren Sie Muster fehlender Daten:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Multiple Imputation
    Muster analysieren...

Abbildung 3-1
Dialogfeld "Muster analysieren"



- Wählen Sie mindestens zwei Analysevariablen aus. Die Prozedur analysiert Muster fehlender Daten für diese Variablen.

Optionale Einstellungen

Analysegewichtung. Diese Variable enthält Analysegewichtungen (Regression oder Stichprobe). Das Verfahren integriert Analysegewichtungen in Zusammenfassungen fehlender Werte. Fälle mit einer negativen oder nullwertigen Analysegewichtung werden ausgeschlossen.

Ausgabe. Die folgende optionale Ausgabe ist verfügbar:

- **Zusammenfassung der fehlenden Werte** Zeigt ein unterteiltes Kreisdiagramm an, das die Anzahl und die Prozentzahlen der Analysevariablen, Fälle oder einzelne Datenwerte enthält, die über einen oder mehrere fehlende Werte verfügen.
- **Muster fehlender Werte.** Zeigt tabulierte Muster fehlender Werte an. Jedes Muster entspricht einer Gruppe von Fällen mit dem gleichen Muster unvollständiger und vollständiger Daten bei Analysevariablen. Sie können diese Ausgabe verwenden, um zu bestimmen, welche monotone Imputationsmethode für Ihre Daten verwendet werden kann und in welchem Maße Ihre Daten einem monotonen Muster entsprechen. Die Prozedur ordnet Analysevariablen, um ein monotones Muster preiszugeben bzw. anzunähern. Wenn kein nicht monotones Muster nach der Neuordnung existiert, können Sie daraus schließen, dass die Daten ein monotones Muster besitzen, wenn die Analysevariablen als solche geordnet sind.
- **Variablen mit der höchsten Frequenz fehlender Werte.** Zeigt eine Tabelle der Analysevariablen, sortiert nach Prozent der fehlenden Werte in absteigender Reihenfolge, an. Die Tabelle enthält deskriptive Statistiken (Mittelwert und Standardabweichung) für metrische Variablen.

  Sie können die maximale Zahl an anzuzeigenden Variablen und den Mindestprozentsatz fehlender Werte für eine Variable, der dargestellt wird, steuern. Es wird die Menge von Variablen angezeigt, die beiden Kriterien entspricht. Zum Beispiel verlangt das Einstellen der Maximalzahl von Variablen auf 50 und des Mindestprozentsatzes fehlender Werte auf 25, dass die Tabelle bis zu 50 Variablen anzeigt, die mindestens 25 % fehlende Werte besitzen. Wenn es 60 Analysevariablen gibt, aber nur 15 25 % oder mehr fehlende Werte haben, enthält die Ausgabe nur 15 Variablen.

Fehlende Datenwerte ersetzen

“Fehlende Datenwerte ersetzen” wird verwendet, um multiple Imputationen zu erzeugen. Die vollständigen Daten-Sets können mit Prozeduren analysiert werden, die Daten-Sets mit multipler Imputation unterstützen. Informationen zur Analyse von Datensets der multiplen Imputation und eine Liste der Verfahren, die diese Daten unterstützen, finden Sie unter Analysieren von Daten multipler Imputation auf S. 29.

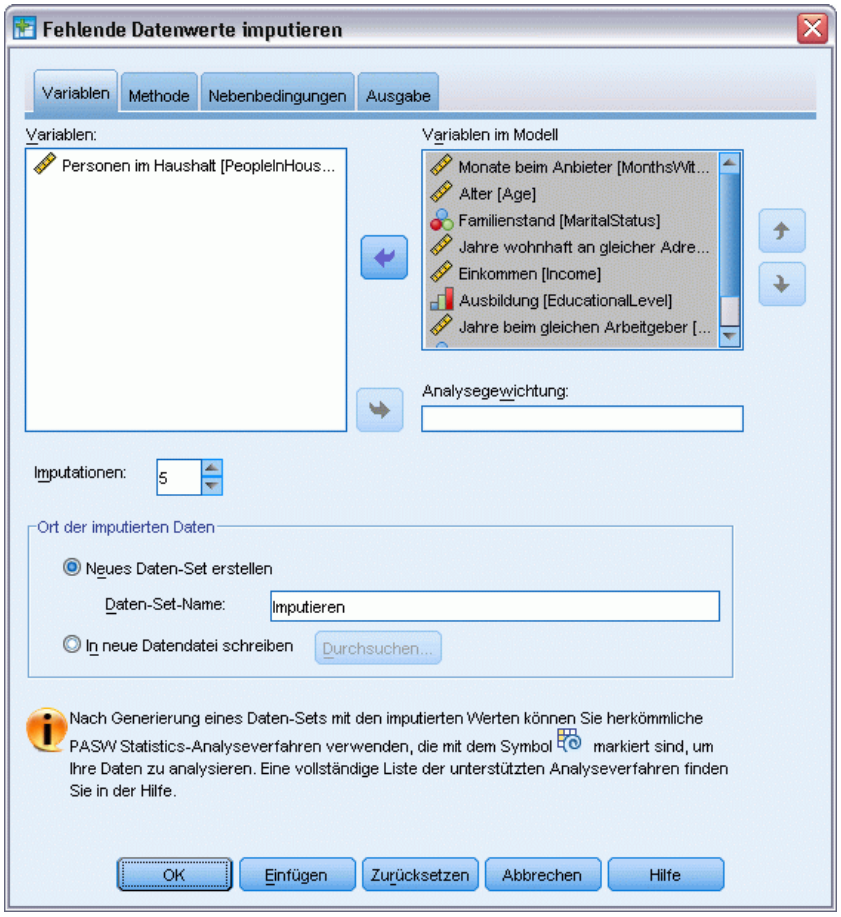
Beispiel. Ein Telekommunikationsanbieter möchte einen besseren Einblick in die Servicenutzungsmuster in seiner Kundendatenbank gewinnen. Er verfügt über die vollständigen Daten der von seinen Kunden genutzten Services, jedoch fehlen in den demographischen Informationen, die das Unternehmen gesammelt hat, einige Werte. Zudem fehlen diese Werte nicht völlig zufällig, daher wird das Daten-Set mithilfe multipler Imputation vervollständigt. Für weitere Informationen siehe Verwendung von multipler Imputation für die Vervollständigung und Analyse einer Daten-Sets in Kapitel 5 auf S. 50.

So ersetzen Sie fehlende Datenwerte:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Multiple Imputation
    Fehlende Datenwerte imputieren...

Abbildung 3-2
Registerkarte "Fehlende Datenwerte ersetzen - Variablen"



- Wählen Sie mindestens zwei Variablen im Imputationsmodell aus. Die Prozedur imputiert mehrere Werte für fehlende Daten für diese Variablen.
- Die Anzahl der zu berechnenden Imputationen. Standardmäßig ist dieser Wert 5.
- Geben Sie ein Daten-Set oder eine Datendatei im PASW Statistics-Format an, in das die imputierten Daten geschrieben werden sollen.

  Das Ausgabe-Daten-Set besteht aus den Originaldaten mit fehlenden Daten plus einem Set von Fällen mit imputierten Werten für jede Imputation. Wenn beispielsweise das ursprüngliche Daten-Set 100 Fälle enthält und Sie haben fünf Imputationen, umfasst das Ausgabe-Daten-Set 600 Fälle. Alle Variablen im Eingabe-Daten-Set sind im Ausgabe-Daten-Set enthalten. Wörterbucheigenschaften (Namen, Labels etc.) von bestehenden Variablen werden in das neue

Daten-Set kopiert. Die Datei enthält auch eine neue Variable, *Imputation_*, eine numerische Variable, die die Imputation angibt (0 für Originaldaten, 1..*n* für Fälle mit imputierten Werten).

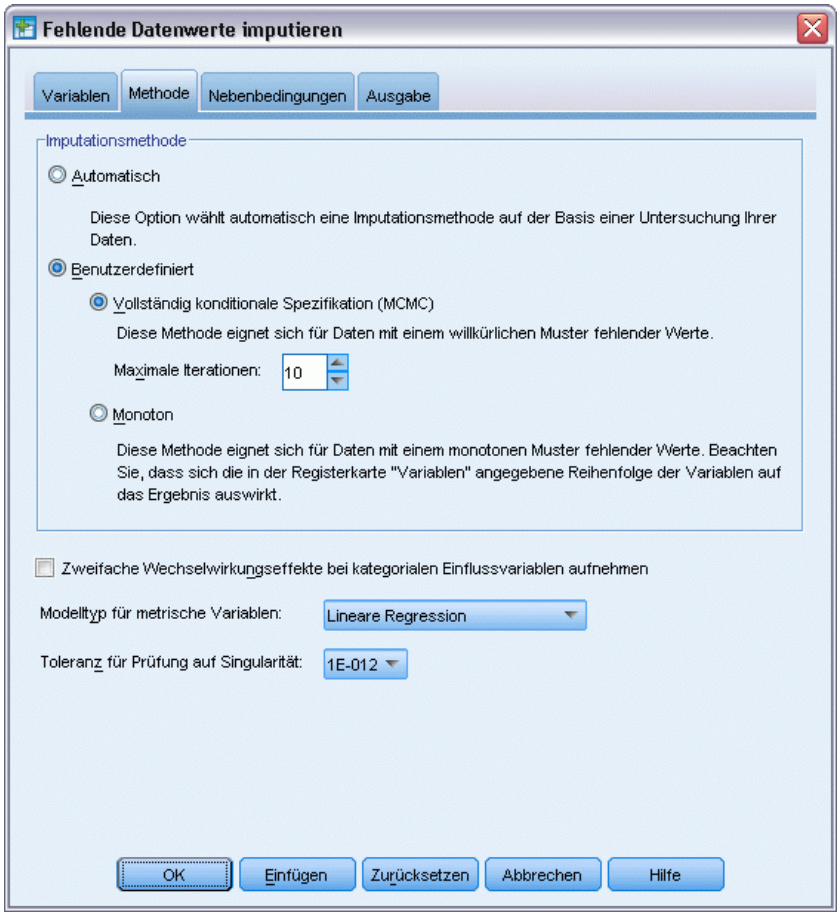
Die Prozedur definiert automatisch die Variable *Imputation_* als aufgeteilte Variable , wenn das Ausgabe-Daten-Set erstellt wird. Wenn bei Ausführung der Prozedur Aufteilungen wirksam sind, enthält das Ausgabe-Daten-Set ein Set an Imputationen für jede Kombination von Werten von ausgeteilten Variablen.

Optionale Einstellungen

Analysegewichtung. Diese Variable enthält Analysegewichtungen (Regression oder Stichprobe). Die Prozedur umfasst Analysegewichtungen in Regressions- und Klassifizierungsmodellen, die verwendet werden, um fehlende Werte zu imputieren. Analysegewichtungen werden auch in Zusammenfassungen imputierter Werte verwendet, zum Beispiel Mittelwert, Standardabweichung und Standardfehler. Fälle mit einer negativen oder nullwertigen Analysegewichtung werden ausgeschlossen.

Methode

Abbildung 3-3
Registerkarte "Fehlende Datenwerte ersetzen - Methode"



Die Registerkarte “Methode” gibt an, wie fehlende Werte einschließlich der verwendeten Modelltypen imputiert werden. Kategoriale Einflussvariablen sind als Indicator (Dummy) kodiert.

Imputationsmethode. Die Methode Automatisch scannt die Daten und verwendet die monotone Methode, wenn die Daten ein monotones Muster fehlender Werte zeigen. Anderenfalls wird die vollständig konditionale Spezifikation verwendet. Wenn Sie sich sicher sind, welche Methode Sie verwenden wollen, können Sie sie als eine Methode unter Benutzerdefiniert angeben.

- **Vollständig konditionale Spezifikation.** Dies ist eine iterative Markov Chain Monte Carlo (MCMC) Methode, die verwendet werden kann, wenn das Muster fehlender Daten willkürlich (monoton oder nicht monoton) ist.

  Für jede Iteration und jede Variable in der in der Variablenliste angegebenen Reihenfolge passt die Methode der vollständig konditionalen Spezifikation ein univariates (einzelne abhängige Variable) Modell mit allen anderen Variablen im Modell als Einflussvariablen an und imputiert dann die fehlenden Werte für die anzupassende Variable. Die Methode wir fortgesetzt, bis die maximale Zahl an Iterationen erreicht ist, und die imputierten Werte in der maximalen Iteration werden in das imputierte Daten-Set gespeichert.

  Maximale Anzahl der Iterationen. Gibt die Anzahl der Iterationen oder Schritte an, die die von der Methode der vollständig konditionalen Spezifikation verwendete Markov-Kette durchläuft. Wenn die Methode der vollständig konditionalen Spezifikation automatisch gewählt wurde, verwendet sie die Standardzahl von 10 Iterationen. Wenn Sie die vollständig konditionale Spezifikation explizit wählen, können Sie eine benutzerdefinierte Zahl an Iterationen angeben. Sie müssen ggf. die Anzahl der Iterationen erhöhen, wenn die Markov-Kette nicht konvergiert. Auf der Registerkarte “Ausgabe” können Sie die Iterationsprotokolldaten der vollständig konditionalen Spezifikation speichern und sie als Diagramm ausgeben, um die Konvergenz zu beurteilen.

- **Monoton.** Dies ist eine nicht iterative Methode, die nur verwendet werden kann, wenn die Daten ein monotones Muster fehlender Werte haben. Ein monotones Muster existiert, wenn Sie die Variablen so ordnen können, dass alle vorhergehenden Variablen auch nicht fehlende Werte haben, wenn eine Variable einen nicht fehlenden Wert hat. Wenn Sie dies als benutzerdefinierte Methode angeben, stellen Sie sicher, die Variablen in der Liste in einer Reihenfolge anzugeben, die ein monotones Muster aufweist.

  Für jede Variable in der monotonen Reihenfolge passt die monotone Methode ein univariates (einzelne abhängige Variable) Modell mit allen vorhergehenden Variablen im Modell als Einflussvariablen an und imputiert dann die fehlenden Werte für die anzupassende Variable. Diese imputierten Werte werden in das imputierte Daten-Set gespeichert.

Zweistufige Interaktionen. Wenn die Imputationsmethode automatisch gewählt wird, enthält das Imputationsmodell für jede Variable eine Konstante und Haupteffekte für Einflussvariablen. Wenn eine bestimmte Methode gewählt wird, können Sie optional alle möglichen zweistufigen Interaktionen in die kategorialen Einflussvariablen aufnehmen.

Modellyp für metrische Variablen. Wenn die Imputationsmethode automatisch gewählt wird, wird lineare Regression als univariates Modell für metrische Variablen verwendet. Wenn eine bestimmte Methode gewählt wird, können Sie alternativ Predictive Mean Matching (PMM) als Modell für metrische Variablen wählen. PMM ist eine Variante der linearen Regression,

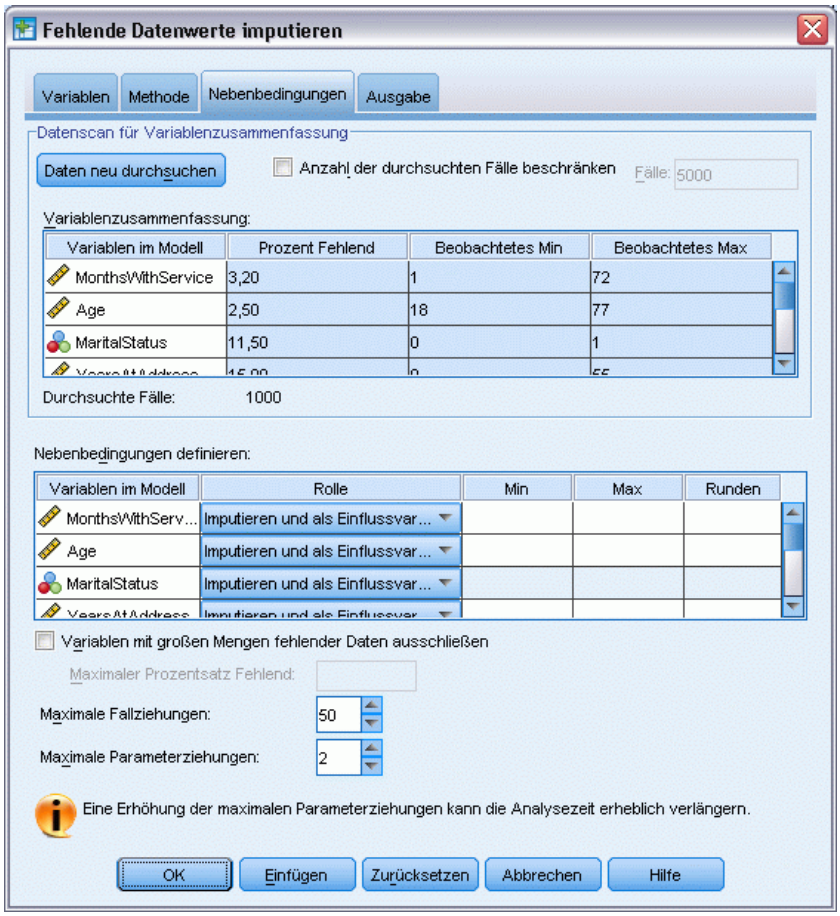
die imputierte Werte, die durch das Regressionsmodell berechnet wurden, mit dem nächsten beobachteten Wert abgleicht.

Logistische Regression wird immer als univariates Modell für kategoriale Variablen verwendet. Unabhängig vom Modelltyp werden kategoriale Einflussvariablen mit Indikatorkodierung (Dummy) gehandhabt.

Toleranz für Prüfung auf Singularität. Singuläre (bzw. nichtinvertierbare) Matrizen weisen linear abhängige Spalten auf, die zu ernsten Problemen für den Schätzalgorithmus führen können. Auch annähernd singuläre Matrizen können zu schlechten Ergebnissen führen, daher behandelt die Prozedur eine Matrix, deren Determinante unter dem Toleranzwert liegt, als singulär. Geben Sie einen positiven Wert ein.

Nebenbedingungen

Abbildung 3-4
Registerkarte "Fehlende Datenwerte ersetzen - Nebenbedingungen"



Mithilfe der Registerkarte "Nebenbedingungen" können Sie die Rolle einer Variablen während der Imputation beschränken und den Bereich der imputierten Werte einer metrischen Variablen so einschränken, dass sie plausibel sind. Zusätzlich können Sie die Analyse auf Variablen mit weniger als einem maximalen Prozentsatz fehlender Werte einschränken.

Daten für Variablenzusammenfassung durchsuchen. Wenn Sie auf Daten durchsuchen klicken, zeigt die Liste Analysevariablen und jeweils den beobachteten Prozentwert für fehlend, Minimum und Maximum. Die Zusammenfassungen können auf allen Fällen oder auf einem Durchlauf der ersten *n* Fälle wie im Textfeld "Fälle" angegeben beruhen. Durch Klicken auf Erneut durchsuchen werden die Verteilungszusammenfassungen aktualisiert.

Nebenbedingungen definieren

- **Rolle.** Hierüber können Sie die Menge der zu imputierenden und/oder als Einflussvariablen zu behandelnden Variablen anpassen. Üblicherweise wird jede Analysevariable im Imputationsmodell sowohl als abhängige Variable als auch als Einflussvariable betrachtet. Die Rolle kann verwendet werden, um die Imputation von Variablen, die Sie Nur als Einflussvariable verwenden wollen, auszuschalten oder um Variablen von der Verwendung als Einflussvariablen (Nur imputieren) auszuschließen und so das Vorhersagemodell kompakter zu machen. Dies ist die einzige Nebenbedingung, die für kategoriale Variablen oder für Variablen, die nur als Einflussvariablen verwendet werden, angegeben werden kann.
- **Min und Max.** In diesen Spalten können Sie die minimal und maximal zulässigen imputierten Werte für metrische Variablen angeben. Wenn ein imputierter Wert außerhalb dieses Bereichs liegt, zieht das Verfahren einen anderen Wert, bis es einen findet, der im Bereich liegt, oder bis die maximale Zahl an Ziehungen erreicht ist (siehe Maximale Ziehungen unten). Diese Spalten sind nur verfügbar, wenn Lineare Regression als Modelltyp für metrische Variablen auf der Registerkarte "Methode" ausgewählt ist.
- **Runden.** Einige Variablen können als metrische Variablen verwendet werden, haben aber Werte, die weiter natürlich beschränkt sein können, z. B. muss die Anzahl der Personen in einem Haushalt eine Ganzzahl sein und der in einem Geschäft ausgegebene Betrag kann keine Bruchteile von Cents umfassen. In dieser Spalte kann die kleinste zulässige Stückelung festgelegt werden. Beispiel: Um ganzzahlige Werte zu erhalten, geben Sie 1 als Rundungswert an; um Werte auf den nächsten Cent zu runden, geben Sie 0,01 an. Im Allgemeinen werden Werte auf das nächste ganzzahlige Vielfache des angegebenen Rundungswerts gerundet. Die folgende Tabelle zeigt, wie sich unterschiedliche Rundungswerte auf den imputierten Wert 6,64823 (vor der Rundung) auswirken.

Rundungswert	Wert, auf den 6,64832 gerundet wird
10	10
1	7
0.25	6.75
0.1	6.6
0.01	6.65

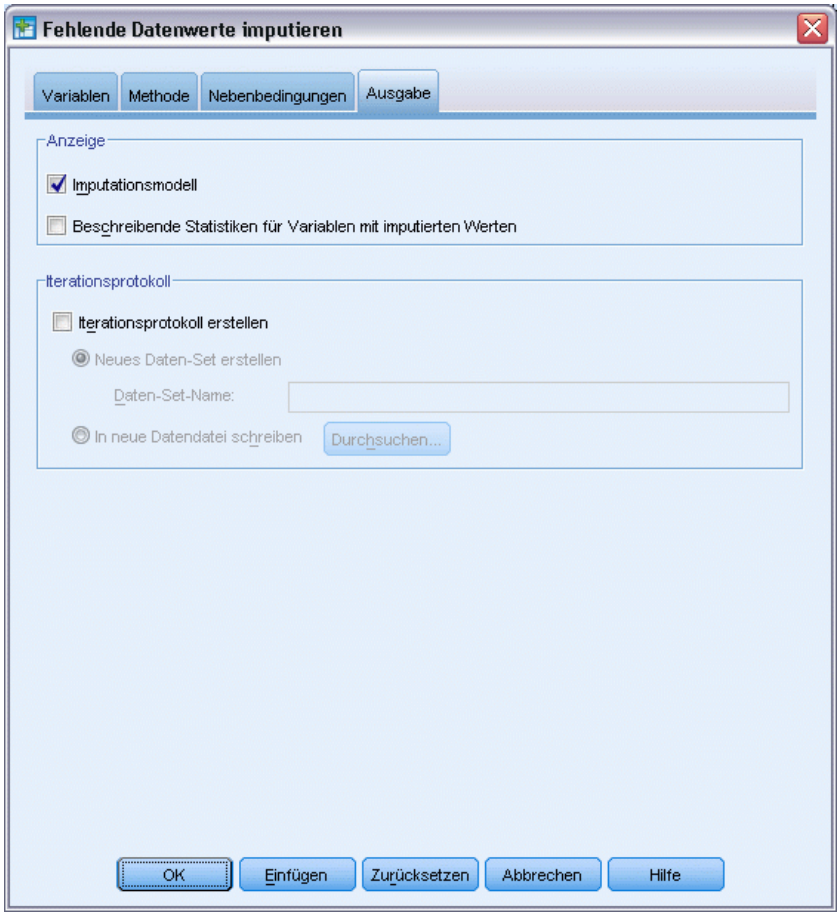
Variablen mit großen Mengen an fehlenden Daten ausschließen. Normalerweise werden Analysevariablen imputiert und als Einflussvariablen verwendet, unabhängig davon, wie viele fehlende Werte sie besitzen, vorausgesetzt, sie haben ausreichend Daten, um ein Imputationsmodell zu schätzen. Sie können Variablen ausschließen, die einen hohen Prozentsatz an fehlenden Werten haben. Wenn Sie zum Beispiel 50 als Maximaler Prozentsatz fehlend angeben, werden Analysevariablen, bei denen mehr als 50 % der Werte fehlen, nicht imputiert und sie werden auch nicht als Einflussvariablen bei Imputationsmodellen verwendet.

Maximale Ziehungen. Wenn Minimum- oder Maximumwerte für imputierte Werte von metrischen Variablen angegeben werden (siehe Min und Max oben), versucht die Prozedur, Werte für einen Fall zu ziehen, bis ein Set an Werten gefunden ist, das innerhalb des angegebenen Bereichs liegt. Wenn mit der angegebenen Zahl an Ziehungen pro Fall kein Set an Werten gefunden wird, zieht die Prozedur ein anderes Set an Modellparametern und wiederholt den Prozess der Fallziehung. Ein Fehler tritt auf, wenn ein Set von Werten im Bereich nicht in der angegebenen Zahl von Fall- und Parameterziehungen gefunden wird.

Beachten Sie, dass höhere Werte eine längere Verarbeitungszeit bedeuten. Wenn die Prozedur lange dauert oder keine geeigneten Ziehungen findet, prüfen Sie die angegebenen Minimum- und Maximumwerte, um sicherzustellen, dass sie angemessen sind.

Ausgabe

Abbildung 3-5
Registerkarte "Fehlende Datenwerte ersetzen - Ausgabe"



Anzeigen. Steuert die Anzeige der Ausgabe. Eine Gesamtimputationszusammenfassung wird immer angezeigt. Sie enthält Tabellen in Bezug auf die Imputationsspezifikationen, die Iterationen (für die Methode vollständiger konditionaler Spezifikation), die abhängigen imputierten Variablen, die abhängigen Variablen, die von der Imputation ausgeschlossen sind, und die

Imputationssequenz. Wenn angegeben, werden auch die Nebenbedingungen für Analysevariablen angezeigt.

- **Imputationsmodell.** Zeigt das Imputationsmodell für abhängige Variablen und Einflussvariablen an und enthält den univariaten Modelltyp, Modelleffekte und die Anzahl der imputierten Werte.
- **Deskriptive Statistik.** Zeigt die deskriptive Statistik für abhängige Variablen an, für die Werte imputiert sind. Für metrische Variablen enthält die deskriptive Statistik Mittelwert, Anzahl, Standardabweichung, Minimum und Maximum für die Original-Eingabedaten (vor der Imputation), imputierte Werte (durch Imputation) und vollständige Daten (Original- und imputierte Werte gemeinsam - durch Imputation). Für kategoriale Variablen enthält die deskriptive Statistik Anzahl und Prozent nach Kategorie für die Original-Eingabedaten (vor der Imputation), imputierte Werte (durch Imputation) und vollständige Daten (Original- und imputierte Werte gemeinsam - durch Imputation).

Iterationsprotokoll. Wenn die Methode vollständiger konditionaler Spezifikation verwendet wird, können Sie ein Daten-Set anfordern, das die Iterationsprotokolldaten für die Imputation nach vollständiger konditionaler Spezifikation enthält. Das Daten-Set enthält Mittelwerte und Standardabweichungen nach Iteration und Imputation für jede metrische abhängige Variable, für die Werte imputiert sind. Sie können die Daten als Diagramm darstellen, um die Beurteilung der Modellkonvergenz zu erleichtern. Für weitere Informationen siehe Prüfen auf FCS-Konvergenz in Kapitel 5 auf S. 69.

Zusätzliche Funktionen beim Befehl MULTIPLE IMPUTATION

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Geben Sie eine Untermenge von Variablen an, für die deskriptive Statistik angezeigt wird (Unterbefehl `IMPUTATIONSUMMARIES`).
- Geben Sie eine Analyse fehlender Muster und Imputation in einem einzigen Lauf der Prozedur an.
- Geben Sie die maximale Anzahl an Modellparametern an, die zulässig sind, wenn eine Variable imputiert wird (Schlüsselwort `MAXMODELPARAM`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Arbeiten mit Daten aus multipler Imputation

Wenn ein Daten-Set multipler Imputation (MI) erstellt wird, wird eine Variable mit dem Namen *Imputation_* und dem Variablenlabel *Imputationszahl* hinzugefügt und das Daten-Set wird danach in aufsteigender Reihenfolge sortiert. Fälle aus dem Original-Daten-Set haben einen Wert von 0. Fälle imputierter Werte sind von 1 bis *M* nummeriert, wobei *M* die Zahl der Imputationen ist.

Wenn Sie ein Daten-Set öffnen, identifiziert das Vorhandensein der *Imputation_* das Daten-Set als mögliches MI-Daten-Set.

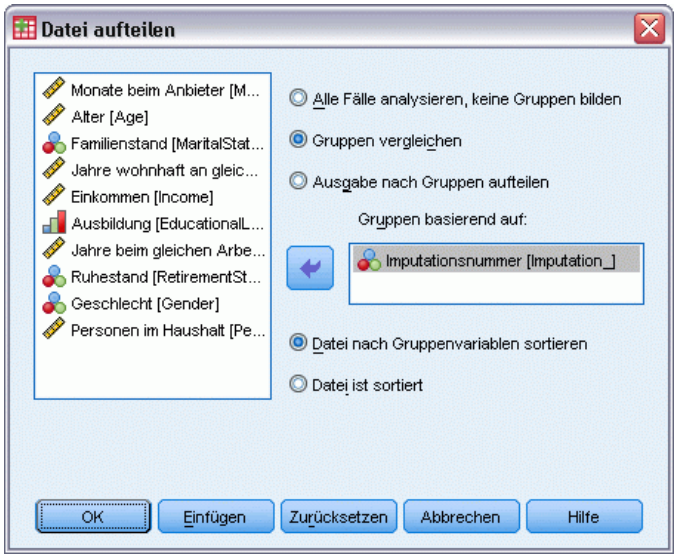
Aktivieren eines Multiple-Imputation-Daten-Sets für die Analyse

Das Daten-Set muss mit der Option Gruppen vergleichen mit *Imputation_* als Gruppierungsvariable aufgeteilt werden, um in Analysen als MI-Daten-Set behandelt zu werden. Sie können auch Aufteilungen bei anderen Variablen definieren.

Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
Datei aufteilen...

Abbildung 3-6
Dialogfeld "Datei aufteilen"



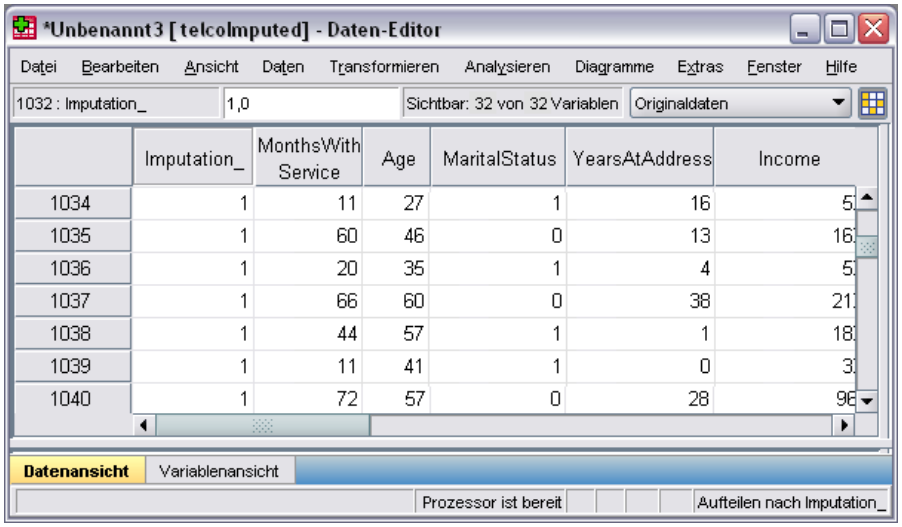
- Wählen Sie die Option Gruppen vergleichen.
- Wählen Sie *Imputationszahl [Imputation_]* als Variable, um Fälle danach zu gruppieren.

Alternativ wird die Datei, wenn Sie Markierungen einschalten (siehe unten), bei *Imputationszahl (Imputation_)* geteilt.

Unterscheidung von imputierten Werten und beobachteten Werten

Sie können imputierte Werte von beobachteten Werten über die Zellenhintergrundfarbe, die Schriftart und den Fettdruck (für imputierte Werte) unterscheiden. Informationen zu den aktivierten Markierungen finden Sie unter Multiple-Imputation-Optionen auf S. 34. Wenn Sie in der aktuellen Sitzung ein neues Daten-Set mit "Fehlende Werte ersetzen" erstellen, werden Markierungen standardmäßig eingeschaltet. Wenn Sie eine gespeicherte Datendatei öffnen, die Imputationen enthält, werden Markierungen ausgeschaltet.

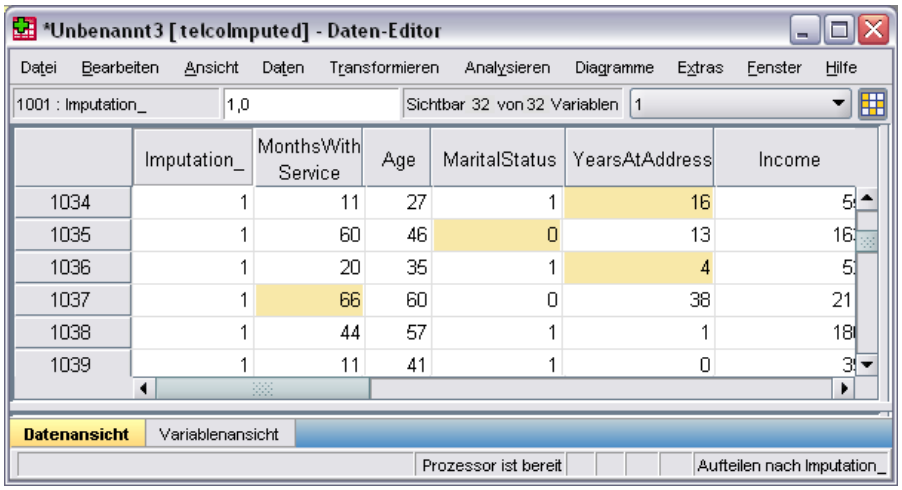
Abbildung 3-7
Daten-Editor mit Imputationsmarkierungen AUS


*Unbenannt3 [telcoImputed] - Daten-Editor
Datei Bearbeiten Ansicht Daten Transformieren Analysieren Diagramme Extras Fenster Hilfe
1032 : Imputation_
1,0
Sichtbar: 32 von 32 Variablen
Originaldaten
Imputation_ MonthsWithService Age MaritalStatus YearsAtAddress Income
1034 1 11 27 1 16 5
1035 1 60 46 0 13 16
1036 1 20 35 1 4 5
1037 1 66 60 0 38 21
1038 1 44 57 1 1 18
1039 1 11 41 1 0 3
1040 1 72 57 0 28 9
Datenansicht Variablenansicht
Prozessor ist bereit
Aufteilen nach Imputation_

Um die Markierungen einzuschalten, wählen Sie aus den Menüs im Daten-Editor:

Ansicht
Imputierte Daten markieren...

Abbildung 3-8
Daten-Editor mit Imputationsmarkierungen EIN


*Unbenannt3 [telcoImputed] - Daten-Editor
Datei Bearbeiten Ansicht Daten Transformieren Analysieren Diagramme Extras Fenster Hilfe
1001 : Imputation_
1,0
Sichtbar 32 von 32 Variablen
1
Imputation_ MonthsWithService Age MaritalStatus YearsAtAddress Income
1034 1 11 27 1 16 5
1035 1 60 46 0 13 16
1036 1 20 35 1 4 5
1037 1 66 60 0 38 21
1038 1 44 57 1 1 18
1039 1 11 41 1 0 3
Datenansicht Variablenansicht
Prozessor ist bereit
Aufteilen nach Imputation_

Alternativ können Sie Markierungen einschalten, indem Sie in der Datenansicht des Daten-Editors auf die Schaltfläche zur Imputationsmarkierung rechts in der Bearbeitungsleiste klicken.

Wechseln zwischen Imputationen

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

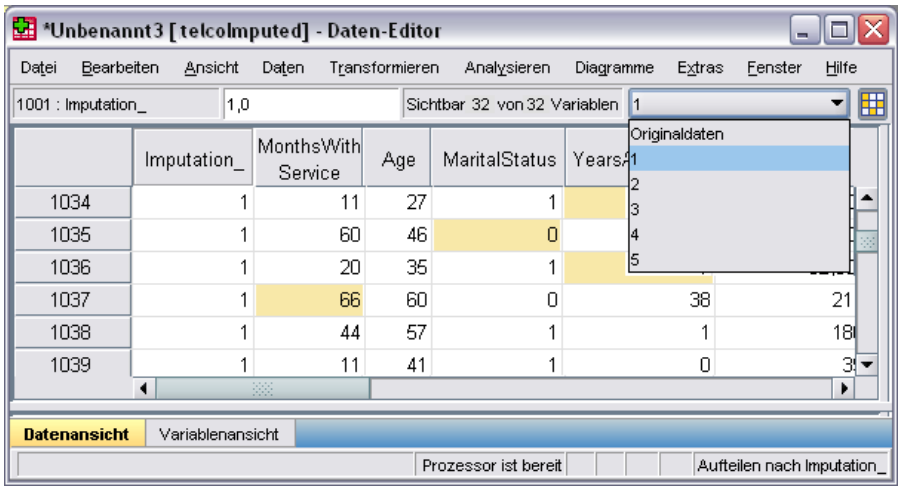
Bearbeiten
Gehe zu Imputation...

- Wählen Sie die Imputation (oder die Originaldaten) aus der Dropdown-Liste.

Abbildung 3-9
Dialogfeld "Gehe zu"

Alternativ können Sie in der Datenansicht des Daten-Editors die Imputation aus der Dropdown-Liste in der Bearbeitenleiste auswählen.

Abbildung 3-10
Daten-Editor mit Imputationsmarkierungen EIN



Die relative Fallposition wird bei der Auswahl der Imputationen beibehalten. Wenn es im Original-Daten-Set 1.000 Fälle gibt, wird Fall 1.034, der 34. Fall in der ersten Imputation, oben im Raster angezeigt. Wenn Sie Imputation 2 in der Dropdown-Liste auswählen, würde Fall 2.034, der 34. Fall in Imputation 2 oben im Raster angezeigt werden. Wenn Sie Originaldaten in der Dropdown-Liste wählen, würde Fall 34 oben im Raster angezeigt werden. Auch die Spaltenposition wird beibehalten, wenn zwischen Imputationen gewechselt wird, sodass der Vergleich von Werten zwischen Imputationen erleichtert wird.

Transformieren und Bearbeiten imputierter Werte

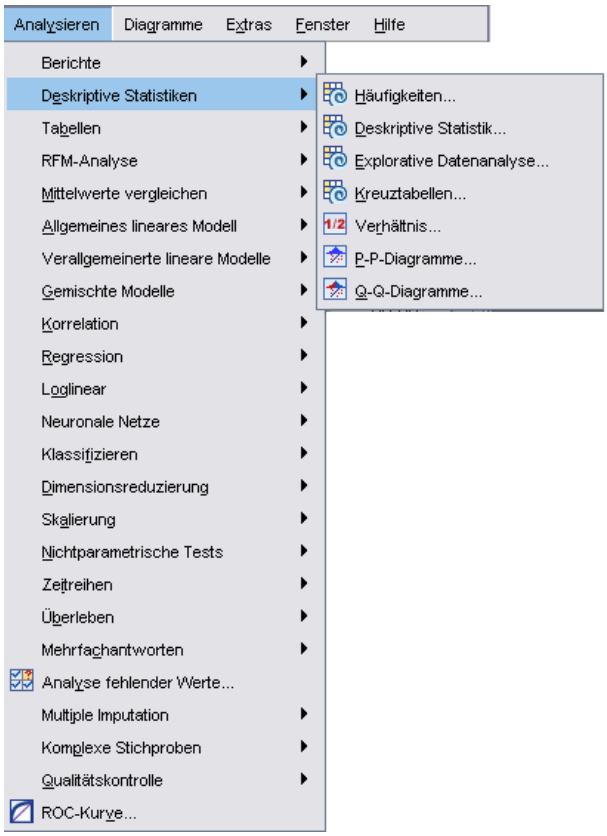
Manchmal müssen Sie Transformationen an imputierten Daten durchführen. Zum Beispiel könnten Sie das Protokoll aller Werte einer Gehaltsvariablen nehmen und das Ergebnis in einer neuen Variablen speichern. Ein Wert, der über imputierte Daten berechnet wurde, wird als imputiert behandelt, wenn er sich von dem Wert, der mit den Originaldaten berechnet wurde, unterscheidet.

Wenn Sie einen imputierten Wert in einer Zelle des Daten-Editors bearbeiten, wird die Zelle immer noch als imputiert behandelt. Es wird nicht empfohlen, imputierte Werte auf diese Weise zu bearbeiten.

Analysieren von Daten multipler Imputation

Viele Prozeduren unterstützen das Pooling von Ergebnissen aus Analysen multipler imputierter Daten-Sets. Wenn Imputationsmarkierungen eingeschaltet sind, wird ein Spezialsymbol neben den Prozeduren angezeigt, die Pooling unterstützen. Im Untermenü “Deskriptive Statistik” des Menüs “Analysieren” zum Beispiel unterstützen “Häufigkeiten”, “Deskriptive Statistik”, “Explorative Datenanalyse” und “Kreuztabellen” Pooling, während “Verhältnisskala”, “P-P-Diagramme” und “Q-Q-Diagramme” kein Pooling unterstützen.

Abbildung 3-11
Menü “Analysieren” mit Imputationsmarkierungen EIN



Sowohl die Tabellenausgabe als auch Modell-PMML unterstützen Pooling. Es gibt keine neue Prozedur für die Anforderung gepoolter Ausgabe. Stattdessen haben Sie über eine neue Registerkarte im Dialogfeld "Optionen" die Möglichkeit, die Ausgabe multipler Imputation zu steuern.

- **Pooling der Tabellenausgabe.** Standardmäßig werden die Ergebnisse, wenn Sie eine unterstützte Prozedur an einem Multiple-Imputation- (MI) Daten-Set ausführen, automatisch für jede Imputation, die Originaldaten (nicht imputiert) und gepoolte (final) Ergebnisse erzeugt, die die Variation über die Imputationen berücksichtigen. Die gepoolten Statistiken unterscheiden sich je nach Prozedur.
- **Pooling von PMML.** Sie können auch gepoolte PMML von unterstützten Prozeduren erhalten, die PMML exportieren. Gepooltes PMML wird auf die gleiche Weise angefordert und wird statt nicht gepoolter PMML gespeichert.

Nicht unterstützte Prozeduren erzeugen entweder gepoolte Ausgabe oder gepoolte PMML-Dateien.

Pooling-Stufen

Die Ausgabe wird mittels einer von zwei Stufen gepoolt:

- **Naive Kombination.** Nur der gepoolte Parameter ist verfügbar.
- **Univariate Kombination.** Der gepoolte Parameter, sein Standardfehler, die Teststatistik und die effektiven Freiheitsgrade, der *p*-Wert, das Konfidenzintervall und die Pooling-Diagnose (Bruchteil der fehlenden Informationen, relative Effizienz, relativer Anstieg der Varianz) werden, wenn verfügbar, angezeigt.

Koeffizienten (Regression und Korrelation), Mittelwerte (und mittlere Differenzen) und Häufigkeiten werden typischerweise in Pools zusammengefasst. Wenn der Standardfehler der Statistik verfügbar ist, wird das univariate Pooling verwendet, andernfalls das naive Pooling.

Prozeduren, die Pooling unterstützen

Die folgenden Prozeduren unterstützen MI-Daten-Sets mit den für jeden Ausgabeteil angegebenen Poolingstufen.

Häufigkeiten

- Die Statistik-Tabelle unterstützt Mittelwerte bei univariatem Pooling (wenn auch der Standardfehler des Mittelwerts angefordert wird) und Gültiges-N und Fehlendes-N bei naivem Pooling.
- Die Tabelle "Häufigkeiten" unterstützt Häufigkeit bei naivem Pooling.

Deskriptive Statistik

- Die Tabelle "Deskriptive Statistiken" unterstützt Mittelwerte bei univariatem Pooling (wenn auch der Standardfehler des Mittelwerts angefordert wird) und N bei naivem Pooling.

Kreuztabellen

- Die Tabelle "Kreuztabelle" unterstützt Anzahl bei naivem Pooling.

Mittelwerte

- Die Tabelle “Bericht” unterstützt Mittelwerte bei univariatem Pooling (wenn auch der Standardfehler des Mittelwerts angefordert wird) und N bei naivem Pooling.

T-Test bei einer Stichprobe

- Die Tabelle “Statistik” unterstützt Mittelwert bei univariatem Pooling und N bei naivem Pooling.
- Die Tabelle “Test” unterstützt Mittelwertdifferenz bei naivem Pooling.

T-Test bei unabhängigen Stichproben

- Die Tabelle “Gruppenstatistik” unterstützt Mittelwert bei univariatem Pooling und N bei naivem Pooling.
- Die Tabelle “Test” unterstützt mittlere Differenz bei univariatem Pooling.

T-Test bei gepaarten Stichproben

- Die Tabelle “Statistik” unterstützt Mittelwerte bei univariatem Pooling und N bei naivem Pooling.
- Die Tabelle “Korrelationen” unterstützt Korrelationen und N bei naivem Pooling.
- Die Tabelle “Test” unterstützt Mittelwert bei univariatem Pooling.

Einfaktorielle ANOVA

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert bei univariatem Pooling und N bei naivem Pooling.
- Die Tabelle “Kontrasttests” unterstützt Kontrastwert bei univariatem Pooling.

GLM Univariate, GLM Multivariate und GLM Wiederholt

- Die Tabelle “Zwischensubjektfaktoren” unterstützt N bei naivem Pooling.
- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Parameterschätzer” unterstützt den Koeffizienten B bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Schätzungen” unterstützt Mittelwert bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Paarweise Vergleiche” unterstützt mittlere Differenz bei univariatem Pooling.

Lineare gemischte Modelle

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Schätzungen fester Effekte” unterstützt Schätzer bei univariatem Pooling.
- Die Tabelle “Schätzungen von Kovarianzparametern” unterstützt Schätzer bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Schätzungen” unterstützt Mittelwert bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Paarweise Vergleiche” unterstützt mittlere Differenz bei univariatem Pooling.

Verallgemeinerte lineare Modelle und verallgemeinerte Schätzungsgleichungen. Diese Prozeduren unterstützen gepooltes PMML.

- Die Tabelle “Informationen zu kategorialen Variablen” unterstützt N und Prozente bei naivem Pooling.
- Die Tabelle “Informationen zu stetigen Variablen” unterstützt N und Mittelwert bei naivem Pooling.
- Die Tabelle “Parameterschätzer” unterstützt den Koeffizienten B bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Schätzkoeffizienten” unterstützt Mittelwert bei naivem Pooling.
- Die Tabelle “Geschätzte Randmittel: Schätzungen” unterstützt Mittelwert bei univariatem Pooling.
- Die Tabelle “Geschätzte Randmittel: Paarweise Vergleiche” unterstützt mittlere Differenz bei univariatem Pooling.

Bivariate Korrelationen

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Korrelationen” unterstützt Korrelationen und N bei naivem Pooling.

Partielle Korrelationen

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Korrelationen” unterstützt Korrelationen bei naivem Pooling.

Lineare Regression. Diese Prozedur unterstützt gepooltes PMML.

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Korrelationen” unterstützt Korrelationen und N bei naivem Pooling.
- Die Tabelle “Koeffizienten” unterstützt B bei univariatem Pooling und Korrelationen bei naivem Pooling.
- Die Tabelle “Korrelationskoeffizienten” unterstützt Korrelationen bei naivem Pooling.
- Die Tabelle “Residuenstatistik” unterstützt Mittelwert und N bei naivem Pooling.

Binäre logistische Regression. Diese Prozedur unterstützt gepooltes PMML.

- Die Tabelle “Variablen in der Gleichung” unterstützt B bei univariatem Pooling.

Multinomiale logistische Regression. Diese Prozedur unterstützt gepooltes PMML.

- Die Tabelle “Parameterschätzer” unterstützt den Koeffizienten B bei univariatem Pooling.

Ordinale Regression

- Die Tabelle “Parameterschätzer” unterstützt den Koeffizienten B bei univariatem Pooling.

Diskriminanzanalyse. Diese Prozedur unterstützt gepooltes Modell-XML.

- Die Tabelle “Gruppenstatistik” unterstützt Mittelwert und Gültiges N bei naivem Pooling.
- Die Tabelle “Gepoolt innerhalb von Gruppenmatrizen” unterstützt Korrelationen bei naivem Pooling.

- Die Tabelle “Kanonische Diskriminanzfunktionskoeffizienten” unterstützt nicht standardisierte Koeffizienten bei naivem Pooling.
- Die Tabelle “Funktionen bei Gruppen-Mittelpunkten” unterstützt nicht standardisierte Koeffizienten bei naivem Pooling.
- Die Tabelle “Klassifizierungsfunktionskoeffizienten” unterstützt Koeffizienten bei naivem Pooling.

Chi-Quadrat-Test

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Häufigkeiten” unterstützt Beobachtetes N bei naivem Pooling.

Test auf Binomialverteilung

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.
- Die Tabelle “Test” unterstützt N, beobachteter Anteil und Testanteil bei naivem Pooling.

Sequenzentest

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.

Kolmogorov-Smirnov-Test bei einer Stichprobe

- Die Tabelle “Deskriptive Statistik” unterstützt Mittelwert und N bei naivem Pooling.

Tests bei zwei unabhängigen Stichproben

- Die Tabelle “Ränge” unterstützt mittlerer Rang und N bei naivem Pooling.
- Die Tabelle “Häufigkeiten” unterstützt N bei naivem Pooling.

Tests bei mehreren unabhängigen Stichproben

- Die Tabelle “Ränge” unterstützt mittlerer Rang und N bei naivem Pooling.
- Die Tabelle “Häufigkeiten” unterstützt Anzahlen bei naivem Pooling.

Tests bei zwei verbundenen Stichproben

- Die Tabelle “Ränge” unterstützt mittlerer Rang und N bei naivem Pooling.
- Die Tabelle “Häufigkeiten” unterstützt N bei naivem Pooling.

Tests bei mehreren verbundenen Stichproben

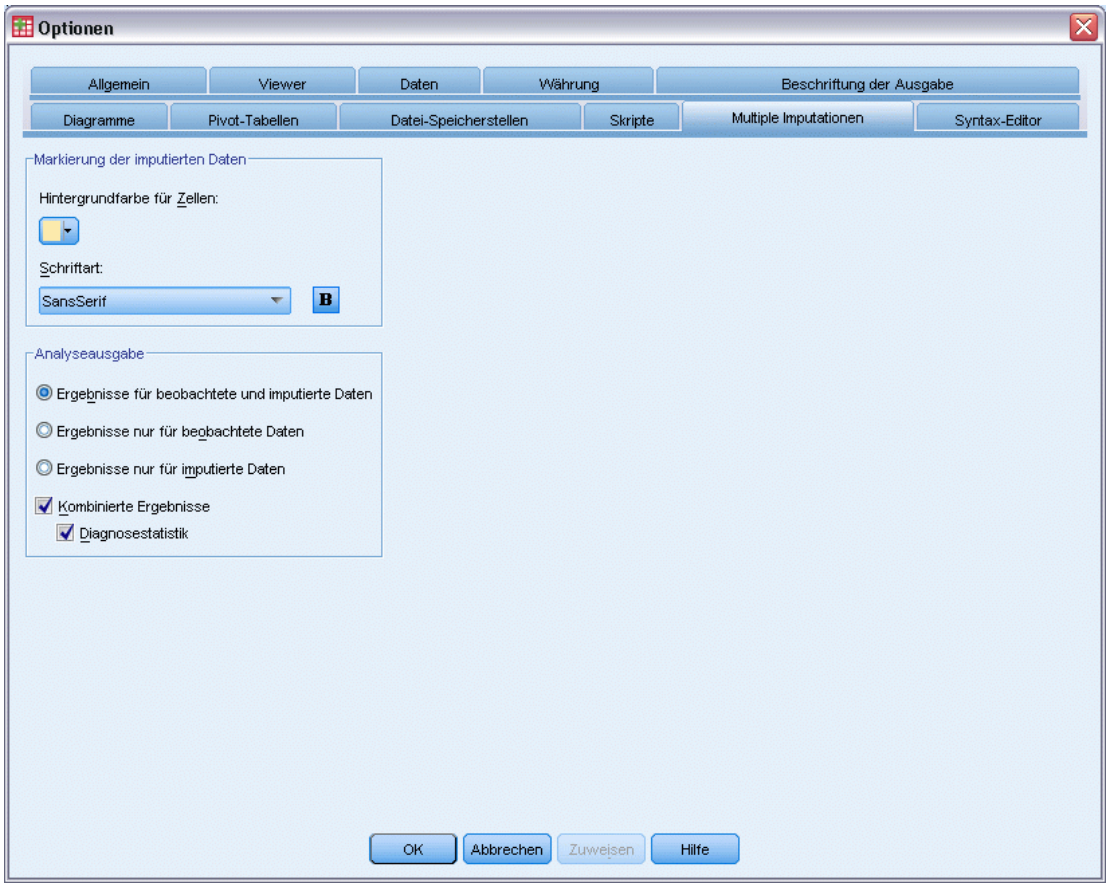
- Die Tabelle “Ränge” unterstützt mittlerer Rang bei naivem Pooling.

Cox-Regression. Diese Prozedur unterstützt gepooltes PMML.

- Die Tabelle “Variablen in der Gleichung” unterstützt B bei univariatem Pooling.
- Die Tabelle “Kovariate Mittelwerte” unterstützt Mittelwert bei naivem Pooling.

Multiple-Imputation-Optionen

Abbildung 3-12
Dialogfeld "Optionen": Registerkarte "Multiple Imputationen"



Die Registerkarte "Multiple Imputationen" steuert zwei Arten von Voreinstellungen für multiple Imputationen:

Erscheinungsbild imputierter Daten. Standardmäßig werden Zellen mit imputierten Daten mit einer anderen Hintergrundfarbe als Zellen mit nicht imputierten Daten angezeigt. Das Erscheinungsbild der imputierten Daten sollte es Ihnen erleichtern, durch ein Daten-Set zu blättern und diese Zellen zu finden. Sie können die Standard-Hintergrundfarbe für die Zellen und die Schriftfamilie ändern und imputierte Daten fett darstellen.

Analyseausgabe. Diese Gruppe steuert die Art der Viewer-Ausgabe, die erzeugt wird, wenn ein multiples imputiertes Daten-Set analysiert wird. Standardmäßig wird die Ausgabe für das Original-Daten-Set (vor der Imputation) und für jedes der imputierten Daten-Sets erzeugt. Zusätzlich werden finale gemeinsame Ergebnisse für die Verfahren erzeugt, die das Pooling von imputierten Daten unterstützen. Bei univariatem Pooling wird auch die Pooling-Diagnose angezeigt. Sie können die Ausgaben, die Sie nicht sehen möchten, jedoch unterdrücken.

So stellen Sie die Optionen für multiple Imputation ein:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Bearbeiten
  Optionen

Klicken Sie auf die Registerkarte “Multiple Imputation”.

Teil II: Beispiele

Kapitel
4

Analyse fehlender Werte

Beschreiben des Musters fehlender Daten

Ein Telekommunikationsanbieter möchte einen besseren Einblick in die Servicenutzungsmustern in seiner Kundendatenbank gewinnen. Das Unternehmen möchte sicherstellen, dass die Daten völlig zufällig fehlen, bevor weitere Analysen durchgeführt werden.

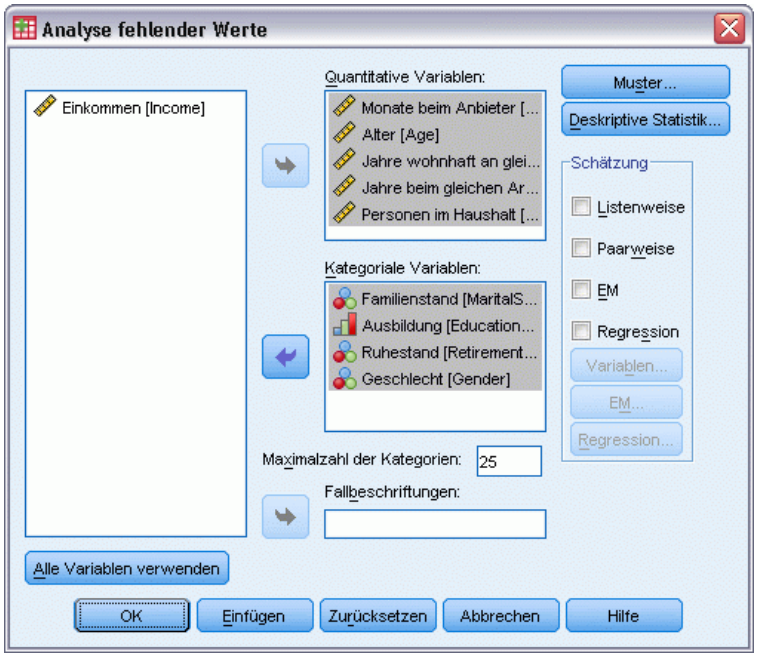
Eine Zufallsstichprobe aus der Kundendatenbank finden Sie in *telco_missing.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Missing Values (Fehlende Werte) 18*.

Durchführen der Analyse zur Anzeige deskriptiver Statistiken

- Zum Ausführen der Prozedur “Analyse fehlender Werte” wählen Sie die folgenden Menübefehle aus:

Analysieren
  Analyse fehlender Werte...

Abbildung 4-1
Dialogfeld "Analyse fehlender Werte"

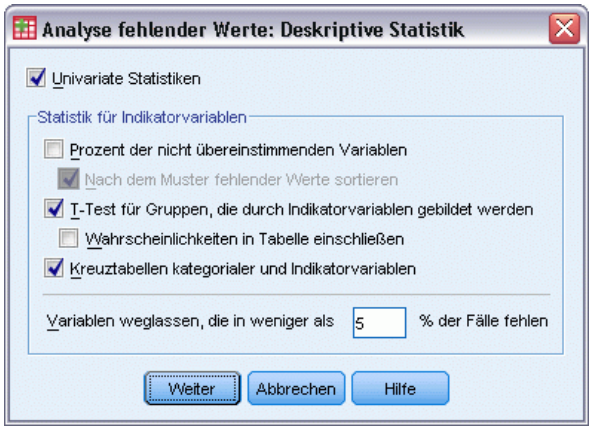


- Wählen Sie *Marital status [marital]* (Familienstand), *Level of education [ed]* (Bildungsniveau), *Retired [retire]* (Im Ruhestand) und *Gender [gender]* (Geschlecht) als kategoriale Variablen aus.
- Wählen Sie *Months with service [tenure]* (Beschäftigungsdauer) bis *Number of people in household [reside]* (Haushaltsgröße) als quantitative (metrische) Variable aus.

Nun könnten Sie die Prozedur durchführen und univariate Statistiken erstellen lassen, wir möchten jedoch zusätzliche deskriptive Statistiken auswählen.

- Klicken Sie auf Deskriptive Statistik.

Abbildung 4-2
Analyse fehlender Werte: Dialogfeld "Deskriptive Statistik"



Im Dialogfeld "Deskriptive Statistik" können Sie verschiedene deskriptive Statistiken angeben, die in der Ausgabe angezeigt werden sollen. Mit den standardmäßig aktivierten univariaten Statistiken können Sie das allgemeine Ausmaß der fehlenden Daten ermitteln, die Statistiken mit Indikatorvariablen bieten jedoch mehr Informationen darüber, wie das Muster der fehlenden Daten in einer Variablen die Werte einer anderen Variablen beeinflussen kann.

- Aktivieren Sie die Option T-Test für Gruppen, die durch Indikatorvariablen gebildet werden.
- Aktivieren Sie die Option Kreuztabellen kategorialer und Indikatorvariablen.
- Klicken Sie auf Weiter.
- Klicken Sie im Hauptdialogfeld "Analyse fehlender Werte" auf die Schaltfläche OK.

Evaluieren der deskriptiven Statistiken

In diesem Beispiel beinhalten die Ausgabe folgende Elemente:

- Univariate Statistiken
- Tabelle der *T*-Tests bei unterschiedlicher Varianz, einschließlich der Untergruppenmittelwerte, wenn eine weitere Variable vorliegt oder fehlt
- Tabellen für jede kategoriale Variable, die die Häufigkeiten der fehlenden Daten für die einzelnen Kategorien in Abhängigkeit von den einzelnen quantitativen (metrischen) Variablen anzeigt

Abbildung 4-3
Tabelle für univariate Statistiken

	N	Mittelwert	Standardabweichung	Fehlend		Anzahl der Extremwerte[a]	
				Anzahl	Prozent	Niedrig	Hoch
tenure	1000	35.56	21.268	32	3.2	0	0
age	1000	41.75	12.573	25	2.5	0	0
address	1000	11.47	9.965	150	15.0	0	9
income	1000	71.1462	83.14424	179	17.9	0	71
employ	1000	11.00	10.113	96	9.6	0	15
reside	1000	2.32	1.431	34	3.4	0	33
marital	1000			115	11.5		
ed	1000			35	3.5		
retire	1000			84	8.4		
gender	1000			42	4.2		

a. Anzahl der Fälle außerhalb des Bereichs (Q1 - 1,5*IQR, Q3 + 1,5*IQR).

Die univariaten Statistiken bieten einen ersten Einblick (für jede Variable gesondert) in das Ausmaß der fehlenden Daten. Die Anzahl der nichtfehlenden Werte für die einzelnen Variablen wird in der Spalte *N* und die Anzahl der fehlenden Werte wird in der Spalte *Fehlend Anzahl* angezeigt. In der Spalte *Fehlend Prozent* wird der Prozentsatz der Fälle mit fehlenden Werten angezeigt. Dieser Wert stellt ein gutes Maß für den Vergleich des Ausmaßes der fehlenden Daten zwischen den verschiedenen Variablen dar. *income (Household income in thousands)* (Einkommen) weist die höchste Anzahl von Fällen mit fehlenden Werten (17,9 %) auf, *age (Age in years)* (Alter) die geringste (2,5 %). *Income* (Einkommen) weist auch die höchste Anzahl an Extremwerten auf.

Abbildung 4-4
T-Tests bei unterschiedlicher Varianz

		MonthsWithService	Age	YearsAtAddress	Income	YearsWithEmployer	PeopleInHousehold
address	t	,4	,3	.	3,5	1,4	1,0
	df	202,2	192,5	.	313,6	191,1	199,5
	Anzahl vorhanden	819	832	850	693	766	824
	Anzahl fehlend	149	143	0	128	138	142
	Mittelwert (Vorhanden)	35,68	41,79	11,47	74,0779	11,20	2,34
	Mittelwert (Fehlend)	34,91	41,49	.	55,2734	9,86	2,21
income	t	-5,0	-8,3	-3,9	.	-5,9	3,6
	df	249,5	222,8	191,1	.	203,3	315,2
	Anzahl vorhanden	793	801	693	821	741	792
	Anzahl fehlend	175	174	157	0	163	174
	Mittelwert (Vorhanden)	33,93	40,01	10,67	71,1462	9,91	2,39
	Mittelwert (Fehlend)	42,97	49,73	14,97	.	15,93	2,02
employ	t	-1,0	-,4	-,7	,5	.	-,3
	df	110,5	110,2	97,6	114,9	.	110,9
	Anzahl vorhanden	877	881	766	741	904	874
	Anzahl fehlend	91	94	84	80	0	92
	Mittelwert (Vorhanden)	35,34	41,69	11,37	71,4953	11.00	2.31
	Mittelwert (Fehlend)	37,70	42,27	12,32	67,9125	.	2,37
marital	t	.0	1.8	1.2	-,8	,9	-2,2
	df	148,1	149,5	138,8	121,2	128,3	134,2
	Anzahl vorhanden	856	862	748	728	805	857
	Anzahl fehlend	112	113	102	93	99	109
	Mittelwert (Vorhanden)	35,56	42,00	11,61	70,3887	11,10	2,28
	Mittelwert (Fehlend)	35,57	39,85	10,43	77,0753	10,17	2,61
retire	t	-,6	-,4	-,4	,3	. .	,2
	df	95,4	94,4	84,0	93,2	. .	99,0
	Anzahl vorhanden	888	893	777	751	904	885
	Anzahl fehlend	80	82	73	70	0	81
	Mittelwert (Vorhanden)	35,44	41,70	11,42	71,3356	11,00	2,32
	Mittelwert (Fehlend)	36,89	42,29	11,96	69,1143	.	2.30

Mithilfe der Tabelle "*T*-Tests bei unterschiedlicher Varianz" können Sie Variablen ermitteln, deren Muster fehlender Werte möglicherweise die quantitativen (metrischen) Variablen beeinflusst. Der *T*-Test wird mithilfe einer Indikatorvariablen berechnet, die angibt, ob eine Variable für einen bestimmten Fall vorhanden ist oder fehlt. Die Untergruppenmittelwerte für die Indikatorvariable werden ebenfalls tabellarisch dargestellt. Beachten Sie, dass nur dann eine Indikatorvariable erstellt wird, wenn eine Variable in mindestens 5 % der Fälle fehlende Werte aufweist.

Es hat den Anschein, dass ältere Befragte weniger häufig ihr Einkommensniveau angeben. Wenn *Income* (Einkommen) fehlt, beträgt der Mittelwert für *Age* (Alter) 49,73, im Vergleich zu 40,01, wenn *Income* (Einkommen) vorhanden ist. In der Tat scheint das Fehlen von *income* (Einkommen) die Mittelwerte mehrerer quantitativer (metrischer) Variablen zu beeinflussen. Dies ist ein Hinweis darauf, dass die Daten möglicherweise nicht völlig zufällig fehlen.

Abbildung 4-5
Kreuztabelle für "Marital status [marital]"

			Total	Unverheiratet	Verheiratet	Fehlend
						SysMis
address	Vorhanden	Anzahl	850	390	358	102
		Prozent	85,0	85,5	83,4	88,7
	Fehlend	% SysMis	15,0	14,5	16,6	11,3
income	Vorhanden	Anzahl	821	380	348	93
		Prozent	82,1	83,3	81,1	80,9
	Fehlend	% SysMis	17,9	16,7	18,9	19,1
employ	Vorhanden	Anzahl	904	418	387	99
		Prozent	90,4	91,7	90,2	86,1
	Fehlend	% SysMis	9,6	8,3	9,8	13,9
retire	Vorhanden	Anzahl	916	423	392	101
		Prozent	91,6	92,8	91,4	87,8
	Fehlend	% SysMis	8,4	7,2	8,6	12,2

Die Kreuztabelle kategorialer Variablen gegenüber Indikatorvariablen zeigt ähnliche Informationen an wie die Tabelle "*T*-Tests bei unterschiedlicher Varianz". Es werden erneut Indikatorvariablen erstellt, allerdings werden sie diesmal zur Berechnung der Häufigkeiten in jeder Kategorie für jede einzelne kategoriale Variable verwendet. Anhand dieser Werte können Sie bestimmen, ob zwischen den verschiedenen Kategorien Unterschiede bei den fehlenden Werten vorliegen.

Wenn wir die Tabelle *marital (Marital status)* (Familienstand) betrachten, scheint die Anzahl der fehlenden Werte in den Indikatorvariablen nicht sonderlich stark zwischen den Kategorien von *marital* zu schwanken. Ob eine Person verheiratet ist oder nicht, scheint keine Auswirkungen darauf zu haben, ob Daten für irgendwelche quantitativen (metrischen) Variablen fehlen. So machten beispielsweise unverheiratete Personen in 85,5 % der Fälle Angaben zu *address (Years at current a)ddress* (Wohnhaft an gleicher Adresse (in Jahren)) und verheiratete Personen in 83,4 % der Fälle. Die Differenz ist minimal und wahrscheinlich zufallsbedingt.

Abbildung 4-6
Kreuztabelle "Level of education [ed]"

			Total	Kein High-School-Abschluss	High-School-Abschluss	College-Besuch	College-Abschluss	Post-Undergraduate-Abschluss	Fehlend
									SysMis
YearsAtAddress	Vorhanden	Anzahl	850	163	240	175	186	56	30
		Prozent	85,0	83,2	85,7	88,4	81,9	87.5	85.7
	Fehlend	% SysMis	15,0	16,8	14,3	11,6	18,1	12.5	14.3
Income	Vorhanden	Anzahl	821	155	229	165	193	50	29
		Prozent	82,1	79,1	81,8	83,3	85,0	78.1	82.9
	Fehlend	% SysMis	17,9	20,9	18,2	16,7	15,0	21.9	17.1
YearsWithEmployer	Vorhanden	Anzahl	904	178	254	178	204	60	30
		Prozent	90,4	90,8	90,7	89,9	89,9	93.8	85.7
	Fehlend	% SysMis	9,6	9,2	9,3	10,1	10,1	6.2	14.3
MaritalStatus	Vorhanden	Anzahl	885	193	278	148	184	52	30
		Prozent	88,5	98,5	99,3	74,7	81,1	81.2	85.7
	Fehlend	% SysMis	11,5	1,5	,7	25,3	18,9	18.8	14.3
RetirementStatus	Vorhanden	Anzahl	916	180	259	180	207	60	30
		Prozent	91,6	91,8	92,5	90,9	91,2	93.8	85.7
	Fehlend	% SysMis	8,4	8,2	7,5	9,1	8,8	6.2	14.3

Betrachten wir nun die Kreuztabelle für *ed (Level of education)* (Bildungsniveau). Wenn der Befragte als Bildungsniveau mindestens "Some college" (Einige Semester am College studiert) angab, ist die Wahrscheinlichkeit, dass Angaben für den Familienstand ("MaritalStatus") fehlen, höher. Mindestens 98,5 % der Befragten ohne College-Ausbildung machten Angaben zum Familienstand. Dagegen gaben nur 81,1 % der Personen mit College-Abschluss ("College degree") ihren Familienstand an. Bei Personen, die einige Semester studiert, aber keinen Abschluss haben ("Some College"), liegt der Wert sogar noch niedriger.

Abbildung 4-7
Kreuztabelle für "Retired [retire]"

			Total	Nein	Ja	Fehlend SysMis
address	Vorhanden	Anzahl	850	744	33	73
		Prozent	85,0	85,0	80,5	86,9
	Fehlend	% SysMis	15,0	15,0	19,5	13,1
income	Vorhanden	Anzahl	821	732	19	70
		Prozent	82,1	83,7	46,3	83,3
	Fehlend	% SysMis	17,9	16,3	53,7	16,7
employ	Vorhanden	Anzahl	904	864	40	0
		Prozent	90,4	98,7	97,6	,0
	Fehlend	% SysMis	9,6	1,3	2,4	100,0
marital	Vorhanden	Anzahl	885	777	38	70
		Prozent	88,5	88,8	92,7	83,3
	Fehlend	% SysMis	11,5	11,2	7,3	16,7

Ein deutlicherer Unterschied ist für *retire (Retired)* (Ruhestandsstatus) zu verzeichnen. Personen, die sich im Ruhestand befinden, geben mit wesentlich geringerer Wahrscheinlichkeit ihr Einkommen an als Personen, die noch nicht im Ruhestand sind. Nur 46,3 % der Kunden im Ruhestand gaben ihr Einkommensniveau ("Income") an, während der Prozentsatz der Personen, die sich nicht im Ruhestand befinden und ihr Einkommensniveau angaben, bei 83,7 % lag.

Abbildung 4-8
Kreuztabelle für "Gender [gender]" (Geschlecht)

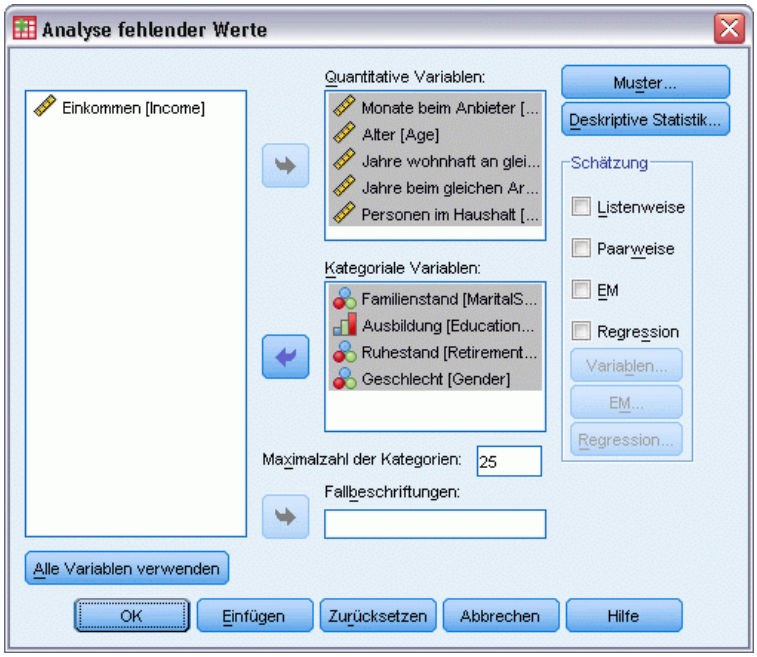
			Total	Männlich	Weiblich	Fehlend SysMis
address	Vorhanden	Anzahl	850	363	456	31
		Prozent	85.0	78.6	91.9	73.8
	Fehlend	% SysMis	15.0	21.4	8.1	26.2
income	Vorhanden	Anzahl	821	381	406	34
		Prozent	82.1	82.5	81.9	81.0
	Fehlend	% SysMis	17.9	17.5	18.1	19.0
employ	Vorhanden	Anzahl	904	412	457	35
		Prozent	90.4	89.2	92.1	83.3
	Fehlend	% SysMis	9.6	10.8	7.9	16.7
marital	Vorhanden	Anzahl	885	400	445	40
		Prozent	88.5	86.6	89.7	95.2
	Fehlend	% SysMis	11.5	13.4	10.3	4.8
retire	Vorhanden	Anzahl	916	420	461	35
		Prozent	91.6	90.9	92.9	83.3
	Fehlend	% SysMis	8.4	9.1	7.1	16.7

Eine weitere Diskrepanz ist für *gender (Gender)* (Geschlecht) offensichtlich. Die Angaben zur Adresse fehlen häufiger bei Männern als bei Frauen. Diese Diskrepanzen könnten zwar zufallsbedingt sein, dies erscheint jedoch unwahrscheinlich. Die Daten scheinen nicht völlig zufällig zu fehlen.

Wir betrachten die Muster der fehlenden Daten, um dies weiter zu untersuchen.

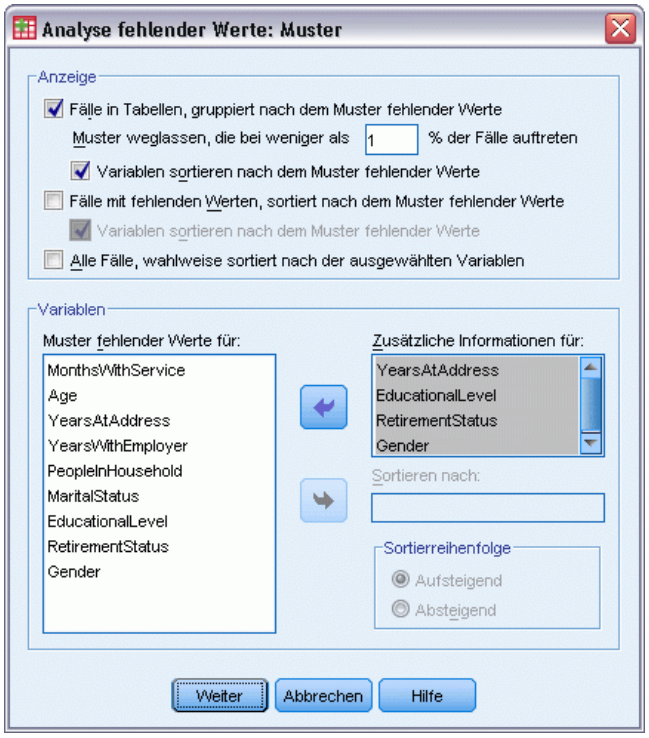
Erneute Durchführung der Analyse zur Anzeige von Mustern.

Abbildung 4-9
Dialogfeld "Analyse fehlender Werte"



- Rufen Sie das Dialogfeld "Analyse fehlender Werte" wieder auf. Das Dialogfeld übernimmt die in der vorherigen Analyse verwendeten Variablen. Ändern Sie dies nicht.
- Klicken Sie auf Muster.

Abbildung 4-10
Dialogfeld "Analyse fehlender Werte: Muster"



Im Dialogfeld "Muster" können Sie verschiedene Mustertabellen auswählen. Wir zeigen Muster in Tabellen, gruppiert nach dem Muster fehlender Werte, an. Da die Muster fehlender Werte in *ed (Level of education)* (Bildungsniveau), *retire (Retired)* (Ruhestandsstatus) und *gender (Gender)* (Geschlecht) Einfluss auf die Daten zu haben schienen, lassen wir weitere Informationen für diese Variablen anzeigen. Außerdem nehmen wir weitere Informationen für *income (Household income in thousands)* (Einkommen) auf, da diese Variable eine so große Anzahl fehlender Werte aufweist.

- Aktivieren Sie die Option Fälle in Tabellen, gruppiert nach dem Muster fehlender Werte.
- Wählen Sie *income* (Einkommen), *ed* (Bildungsniveau), *retire* (Ruhestandsstatus) und *gender* (Geschlecht) aus und fügen Sie sie zur Liste "Zusätzliche Informationen für" hinzu.
- Klicken Sie auf Weiter.
- Klicken Sie im Hauptdialogfeld "Analyse fehlender Werte" auf die Schaltfläche OK.

Evaluieren der Mustertabelle

Abbildung 4-11
Tabelle "Muster in Tabellen"

Anzahl der Fälle	Muster fehlender Werte[a]										Vollständig, wenn ...[b]	Income[c]	EducationalLevel[d]					RetirementStatus[d]		Gender[d]	
	Age	PeopleInHousehold	MonthsWithService	EducationalLevel	Gender	RetirementStatus	YearsWithEmployer	MaritalStatus	YearsAtAddress	Income			Kein High-School-Abschluss	High-School-Abschluss	College-Besuch	College-Abschluss	Post-Undergraduate-Abschluss	Nein	Ja	Männlich	Weiblich
475											475	76,5853	99	157	87	101	31	...	12	201	274
109										X	584	.	27	35	19	17	11	95	14	47	62
16									X	X	687	.	5	9	0	1	1	12	4	12	4
87									X		562	54,4368	21	27	9	24	6	85	2	66	21
13		X									488	56,0000	4	3	2	3	1	13	0	4	9
60								X			535	77,2167	1	2	27	24	6	59	1	35	25
16				X							491	47,8125	0	0	0	0	0	16	0	6	10
17			X								492	76,2353	2	7	3	4	1	17	0	7	10
18					X						493	54,1111	3	7	4	4	0	17	1	0	0
16								X		X	660	.	0	0	7	8	1	14	2	6	10
37						X	X				520	59,4595	9	14	5	8	1	0	0	15	22

Muster mit weniger als 1% Fällen (10 oder weniger) werden nicht angezeigt.

a. Variablen sind nach Mustern fehlender Werte sortiert.

b. Anzahl der vollständigen Fälle, wenn die in diesem Muster fehlenden Variablen (mit X gekennzeichnet) nicht verwendet werden.

c. Mittelwerte bei jedem eindeutigen Muster.

d. Häufigkeitsverteilung bei jedem eindeutigen Muster.

Die Tabelle "Muster in Tabellen" zeigt an, ob die Daten tendenziell für mehrere Variablen in einzelnen Fällen fehlen. Sie können damit also ermitteln, ob die Daten gemeinsam fehlen.

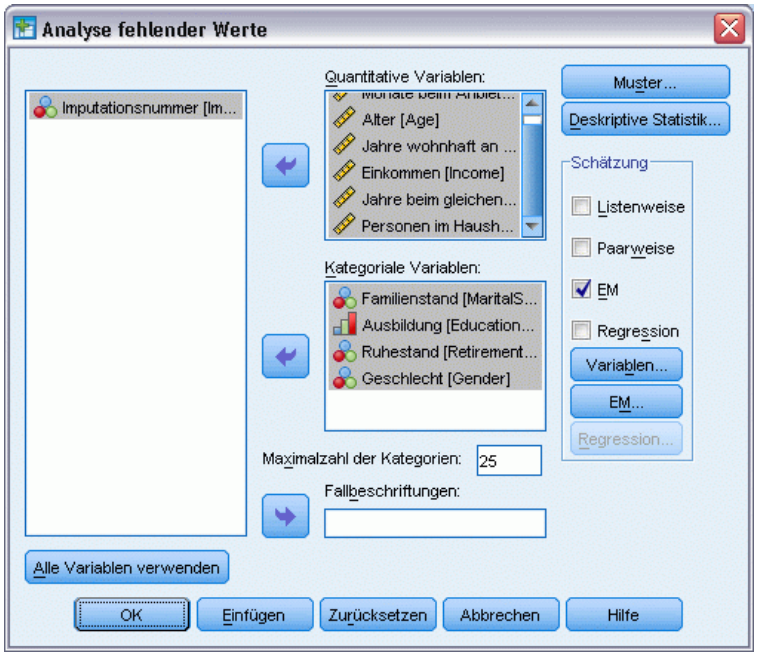
Es gibt drei Muster für gemeinsam fehlende Daten, die in mehr als 1 % der Fälle vorkommen. Die Variablen *employ (Years with current employer)* (Jahre beim derzeitigen Arbeitgeber) und *retire (Retired)* (Ruhestand) fehlen zusammen öfter als die anderen Paare. Dies überrascht nicht, da *retire* (Ruhestand) und *employ* (Jahre beim derzeitigen Arbeitgeber) ähnliche Informationen erfassen. Wenn Sie nicht wissen, ob ein Befragter sich im Ruhestand befindet, wissen Sie vermutlich auch nicht, wie viele Jahre die betreffende Person beim derzeitigen Arbeitgeber beschäftigt ist.

Der Mittelwert für *income (Household income in thousands)* (Einkommen) scheint in Abhängigkeit vom Muster fehlender Werte erheblich zu schwanken. Insbesondere ist der Mittelwert für *Income* (Einkommen) wesentlich höher für die 6 % (60 von 1000) der Fälle, in denen *marital (Marital status)* (Familienstand) fehlt. (Dieser Wert ist auch höher, wenn *tenure (Months with service)* (Beschäftigung) fehlt, doch dieses Muster betrifft nur 1,7 % der Fälle.) Erinnern Sie sich, dass die Personen mit einem höheren Bildungsniveau die Frage nach dem Ehestand weniger häufig beantworteten. Dieser Trend ist in den für *ed (Level of education)* (Bildungsniveau) angezeigten Häufigkeiten zu sehen. Wir könnten den Anstieg bei *income* (Einkommen) möglicherweise erklären, indem wir annehmen, dass die Personen mit einem höheren Bildungsniveau mehr Geld verdienen und weniger häufig ihren Familienstand angeben.

Wenn wir die deskriptiven Statistiken und die Muster fehlender Daten betrachten, können wir möglicherweise folgern, dass die Daten nicht völlig zufällig fehlen. Wir können diese Schlussfolgerung mit dem MCAR-Test nach Little überprüfen, der mit den EM-Schätzern abgedruckt ist.

Erneute Durchführung der Analyse für den MCAR-Test nach Little

Abbildung 4-12
Dialogfeld "Analyse fehlender Werte"



- Rufen Sie das Dialogfeld "Analyse fehlender Werte" wieder auf.
- Klicken Sie auf EM.
- Klicken Sie auf OK.

Abbildung 4-13
Tabelle "Geschätzte Randmittel"

MonthsWithService	Age	Income	YearsWithEmployer	YearsAtAddress	PeopleInHousehold
36,12	41,91	77,3941	11,22	11,58	2,29

a. MCAR-Test nach Little: Chi-Quadrat = 179,836, DF = 107, Sig. = ,000

Die Ergebnisse des MCAR-Tests nach Little werden jeweils in den Fußnoten der Tabellen für EM-geschätzte Statistiken angezeigt. Die Nullhypothese für den MCAR-Test nach Little lautet, dass die Daten in völlig zufälliger Weise fehlen (missing completely at random – MCAR). Daten fehlen völlig zufällig (MCAR), wenn das Muster der fehlenden Werte nicht von den Datenwerten abhängt. Da der Signifikanzwert in unserem Beispiel weniger als 0,05 beträgt, können wir folgern, dass die Daten *nicht* völlig zufällig fehlen. Dies bestätigt die Schlussfolgerung, die wir aus den deskriptiven Statistiken und den Mustern in Tabellen gezogen haben.

Da die Daten nicht völlig zufällig fehlen, ist es an dieser Stelle nicht sicher, Fälle mit fehlenden Werten oder einzeln imputierten fehlenden Werten listenweise zu löschen. Dennoch können Sie Multiple Imputation verwenden, um diese Datenmenge weiter zu analysieren.

Kapitel
5

Multiple Imputation

Verwendung von multipler Imputation für die Vervollständigung und Analyse einer Daten-Sets

Ein Telekommunikationsanbieter möchte einen besseren Einblick in die Servicenutzungsmustern in seiner Kundendatenbank gewinnen. Er verfügt über die vollständigen Daten der von seinen Kunden genutzten Services, jedoch fehlen in den demographischen Informationen, die das Unternehmen gesammelt hat, einige Werte. Zudem fehlen diese Werte nicht völlig zufällig, daher wird das Daten-Set mithilfe multipler Imputation vervollständigt.

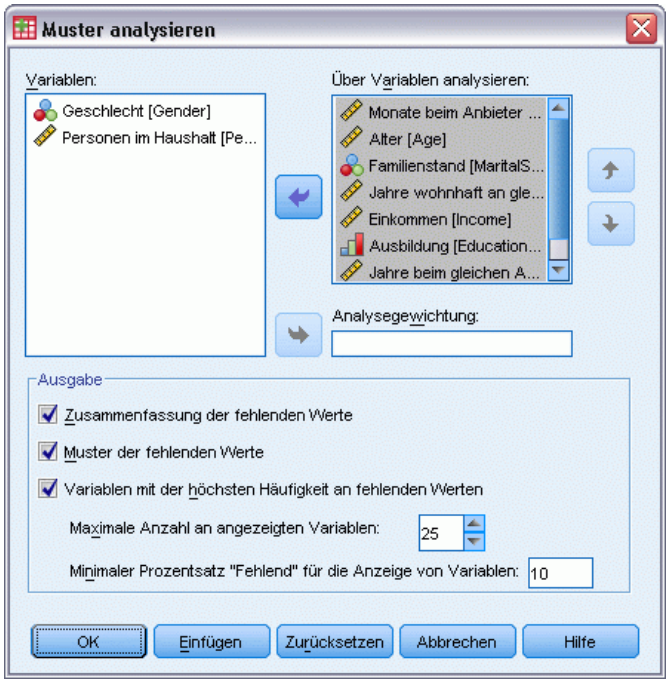
Eine Zufallsstichprobe aus der Kundendatenbank finden Sie in *telco_missing.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A in *PASW® Missing Values (Fehlende Werte) 18*.

Analyse der Muster fehlender Werte

- Sehen Sie sich als ersten Schritt die Muster fehlender Daten an. Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Multiple Imputation
    Muster analysieren...

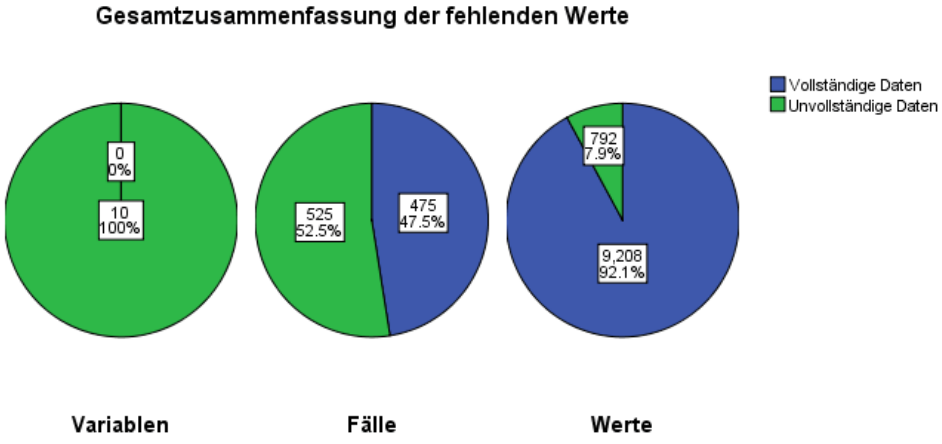
Abbildung 5-1
Muster analysieren, Dialogfeld



- Wählen Sie *Months with service [tenure]* (Beschäftigungsdauer) bis *Number of people in household [reside]* (Haushaltsgröße) als Analysevariable aus.

Gesamtzusammenfassung

Abbildung 5-2
Gesamtzusammenfassung der fehlenden Werte



Die Gesamtzusammenfassung der fehlenden Werte zeigt drei Kreisdiagramme an, die unterschiedliche Aspekte fehlender Werte in den Daten darstellen.

- Das Diagramm *Variablen* zeigt, dass jede der 10 Analysevariablen mindestens einen fehlenden Wert in einem Fall besitzt.
- Das Diagramm *Fälle* zeigt, dass 525 der 1.000 Fälle mindestens einen fehlenden Wert in einer Variable besitzen.
- Das Diagramm *Werte* zeigt, dass 792 der 10.000 Werte (Fälle × Variablen) fehlen.

Jeder Fall mit fehlenden Werten besitzt im Durchschnitt fehlende Werte bei ungefähr 1,5 der 10 Variablen. Ein **listenweiser Ausschluss** würde zu einem Verlust eines Großteils der Informationen in dem Daten-Set führen.

Variablenauswertung

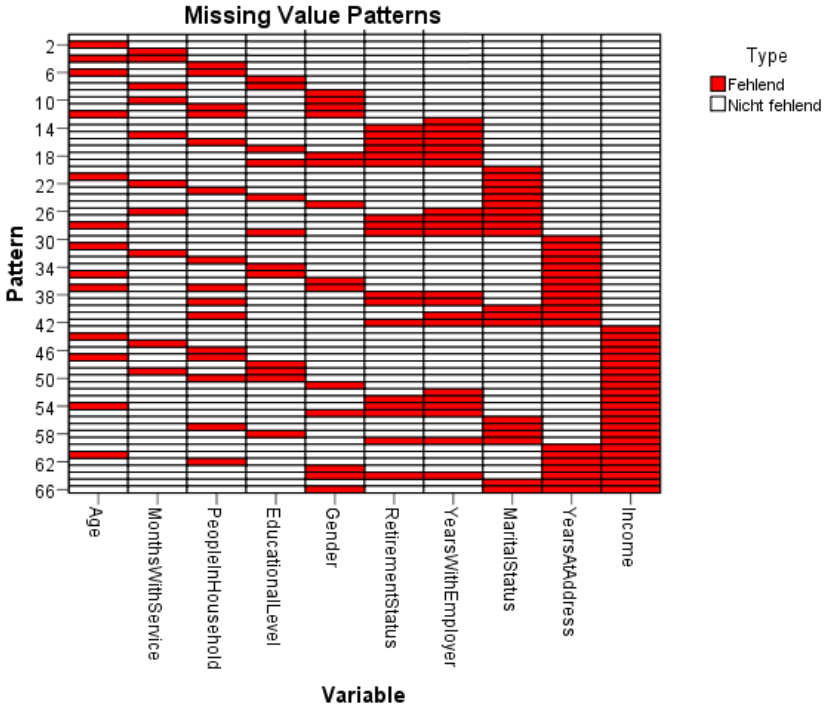
Abbildung 5-3
Variablenauswertung

	Fehlend				
	N	Prozent	Gültige N	Mittelwert	Standardabweichung
Household income in ...	179	17,9%	821	71,1462	83,14424
Years at current address	150	15,0%	850	11,47	9,965
Marital status	115	11,5%	885		

Die Variablenzusammenfassung wird für Variablen mit mindestens 10 % fehlenden Werte angezeigt und zeigt die Anzahl und den Prozentsatz fehlender Werte für jede Variable in der Tabelle. Sie zeigt zudem die mittlere und Standardabweichung für die gültigen Werte der metrischen Variablen und die Anzahl an gültigen Werten für alle Variablen an. *Household income in thousands* (Haushaltseinkommen in Tausend), *Years at current address* (Jahre an der aktuellen Adresse) und *Marital status* (Familienstand) haben die meisten fehlenden Werte in dieser Reihenfolge.

Muster

Abbildung 5-4
Muster fehlender Werte

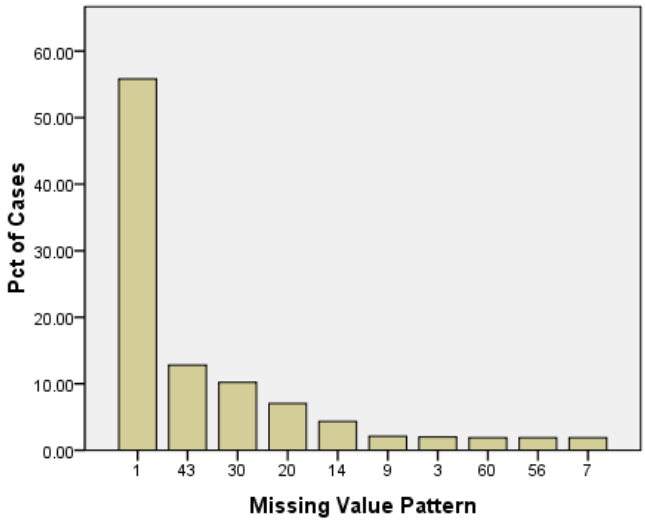


Das Diagramm “Muster” zeigt Muster fehlender Werte für die Analysevariablen an. Jedes Muster entspricht einer Gruppe von Fällen mit dem gleichen Muster unvollständiger und vollständiger Daten. Zum Beispiel stellt Muster 1 Fälle dar, die keine fehlenden Werte besitzen, während Muster 33 Fälle darstellt, die fehlende Werte bei *reside (Number of people in household)* (Haushaltsgröße) und *address (Years at current address)* (Jahre an der aktuellen Adresse) besitzen. Muster 66 stellt Fälle dar, die fehlende Werte bei *gender (Gender)* (Geschlecht), *marital (Marital status)* (Familienstand), *address* (Adresse) und *income (Household income in thousands)* (Einkommen) besitzen. Ein Daten-Set kann potenziell $2^{\text{Anzahl an Variablen}}$ Muster haben. Bei 10 Analysevariablen ist das $2^{10}=1024$. Es werden jedoch nur 66 Muster in den 1.000 Fällen im Daten-Set dargestellt.

Das Diagramm ordnet Analysevariablen und Muster, um Monotonie, falls vorhanden, aufzuzeigen. Speziell werden Variablen von links nach rechts in aufsteigender Reiehnfolge der fehlenden Werte geordnet. Die Muster werden dann zuerst nach der letzten Variable (nicht fehlende Werte zuerst, dann fehlende Werte), dann nach der zweiten bis zur letzten Variable usw. sortiert. Dabei wird von rechts nach links vorgegangen. So wird aufgezeigt, welche monotone Imputationsmethode für Ihre Daten verwendet werden kann und in welchem Maße Ihre Daten einem monotonen Muster entsprechen. Wenn die Daten monoton sind, sind alle fehlenden Zellen und nicht fehlenden Zellen im Diagramm fortlaufend. Es gibt also keine “Inseln” nicht fehlender Zellen im unteren rechten Teil des Diagramms und keine “Inseln” fehlender Zellen im oberen linken Teil des Diagramms.

Dieses Daten-Set ist monoton und es gibt viele Werte, die imputiert werden müssten, um Monotonie zu erreichen.

Abbildung 5-5
Musterhäufigkeiten



Wenn Muster angefordert werden, zeigt ein begleitendes Balkendiagramm den Prozentsatz an Fällen für jedes Muster an. Das zeigt, dass über die Hälfte der Fälle im Daten-Set Muster 1 besitzen. Das Diagramm fehlender Werte zeigt, dass dies das Muster für Fälle ohne fehlende Werte ist. Muster 43 stellt Fälle mit einem fehlenden Wert bei *income*, Muster 30 Fälle mit einem fehlenden Wert bei *address* und Muster 20 Fälle mit einem fehlenden Wert bei *marital* dar. Die große Mehrheit der Fälle, ungefähr 4 von 5, werden durch diese vier Muster dargestellt. Muster 14, 60 und 56 sind die einzigen Muster unter den zehn am häufigsten auftretenden Mustern, um Fälle mit fehlenden Werten bei mehr als einer Variable darzustellen.

Die Analyse fehlender Muster hat keine bestimmten Hindernisse für die multiple Imputation gezeigt, abgesehen davon, dass die Verwendung der monotonen Methode nicht wirklich praktikabel ist.

Automatische Imputation fehlender Werte

Jetzt sind Sie bereit, die Imputation von Werten zu beginnen. Wir beginnen mit einem Durchlauf mit automatischen Einstellungen, bevor wir aber Imputationen anfordern, legen wir den Startwert fest. Durch die Festlegung des Startwerts können sie die Analyse exakt reproduzieren.

- Zur Festlegung des Startwerts wählen Sie die folgenden Menübefehle aus:

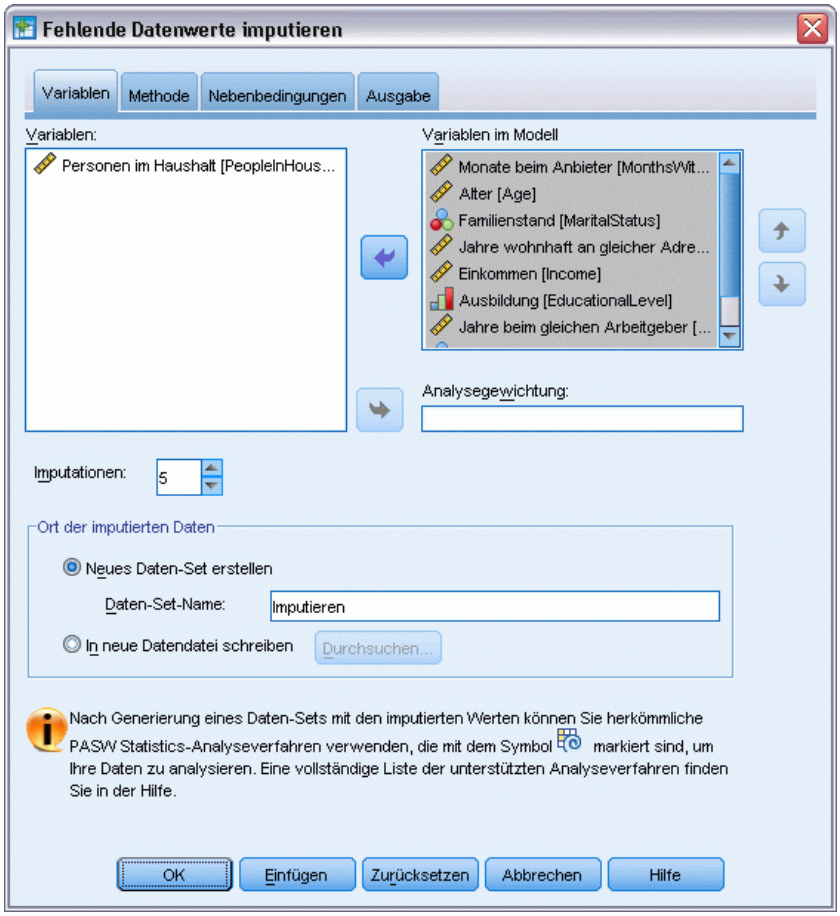
Transformieren
Zufallszahlengeneratoren...

Abbildung 5-6
Dialogfeld "Zufallszahlengenerator"

- Wählen Sie Zufallszahlengenerator bestimmen.
- Wählen Sie Mersenne-Twister.
- Wählen Sie Anfangswert festlegen.
- Wählen Sie Fester Wert und geben Sie 20070525 als Wert ein.
- Klicken Sie auf OK.
- Um mehrfach fehlende Datenwerte zu ersetzen, wählen Sie aus dem Menü:

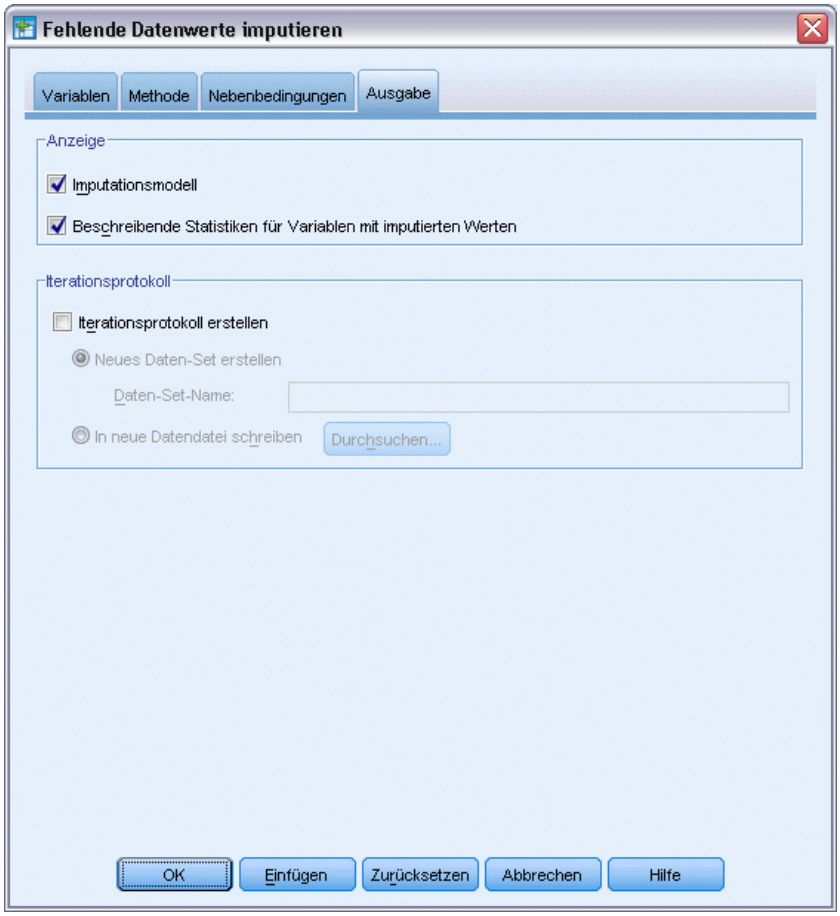
Analysieren
  Multiple Imputation
    Fehlende Datenwerte ersetzen...

Abbildung 5-7
Fehlende Datenwerte ersetzen, Dialogfeld



- Wählen Sie *Months with service [tenure]* (Beschäftigungsdauer) bis *Number of people in household [reside]* (Haushaltsgröße) als Variablen im Imputationsmodell aus.
- Geben Sie telcoImputed als Daten-Set an, in das die imputierten Daten gespeichert werden sollen.
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 5-8
Registerkarte "Ausgabe"



- Wählen Sie Deskriptive Statistik für Variablen mit imputierten Werten.
- Klicken Sie auf OK.

Imputationsspezifikationen

Abbildung 5-9
Imputationsspezifikationen

Imputationsmethode	Automatisch
Anzahl an Imputationen	5
Modell für metrische ...	Lineare Regression
In Modellen enthaltene ...	(ohne)
Maximaler Prozentsatz ...	100,0%

Die Tabelle "Imputationsspezifikationen" gibt eine nützliche Übersicht, mit der Sie sicherstellen können, dass die Angaben richtig waren.

Imputationsergebnisse

Abbildung 5-10
Imputationsergebnisse

Imputationsmethode		Vollständig konditionale Spezifikation
Iterationen der vollständig konditionalen Spezifikationsmethode		10
Abhängige Variablen	Imputiert	MonthsWithService,Age,MaritalStatus, YearsAtAddress,Income,EducationalLevel, YearsWithEmployer,RetirementStatus, Gender,PeopleInHousehold
	Nicht imputiert (zu viele fehlende Werte)	
	Nicht imputiert (keine fehlenden Werte)	
Imputationssequenz		Age,MonthsWithService,PeopleInHousehold, EducationalLevel,Gender,RetirementStatus, YearsWithEmployer,MaritalStatus, YearsAtAddress,Income

Die Imputationsergebnisse geben einen Überblick dessen, was während des Imputationsvorgangs tatsächlich geschieht. Beachten Sie insbesondere Folgendes:

- Die Imputationsmethode in der Spezifikationentabelle war “Automatisch” und die von der automatischen Methodenauswahl gewählten Methode war “Vollständig konditionale Spezifikation”.
- Alle angeforderten Variablen wurden imputiert.
- Die Imputationssequenz ist die Reihenfolge, in der die Variablen auf der *x*-Achse im Diagramm “Muster fehlender Werte” erscheinen.

Imputationsmodelle

Abbildung 5-11
Imputationsmodelle

	Modell			
	Typ	Effekte	Fehlende Werte	Imputierte Werte
Age in years	Lineare Regression	ed,gender,retire,marital,tenure, reside,employ,address,income	25	125
Months with service	Lineare Regression	ed,gender,retire,marital,age, reside,employ,address,income	32	160
Number of people in household	Lineare Regression	ed,gender,retire,marital,age, tenure,employ,address,income	34	170
Level of education	Logistische Regression	gender,retire,marital,age,tenure, reside,employ,address,income	35	175
Gender	Logistische Regression	ed,retire,marital,age,tenure, reside,employ,address,income	42	210
Retired	Logistische Regression	ed,gender,marital,age,tenure, reside,employ,address,income	84	420
Years with current employer	Lineare Regression	ed,gender,retire,marital,age, tenure,reside,address,income	96	480
Marital status	Logistische Regression	ed,gender,retire,age,tenure, reside,employ,address,income	115	575
Years at current address	Lineare Regression	ed,gender,retire,marital,age, tenure,reside,employ,income	150	750
Household income in thousands	Lineare Regression	ed,gender,retire,marital,age, tenure,reside,employ,address	179	895

Die Tabelle “Imputationsmodelle” gibt weitere Details an, wie jede Variable imputiert wurde. Beachten Sie insbesondere Folgendes:

- Die Variablen werden in der Reihenfolge der Imputationssequenz aufgeführt.
- Metrische Variablen werden mit linearer Regression modelliert, kategoriale Variablen mit logistischer Regression.
- Jedes Modell verwendet alle anderen Variablen als Haupteffekte.
- Die Anzahl der fehlenden Werte für jede Variable wird zusammen mit der Gesamtzahl an imputierten Werten für diese Variable (Anzahl fehlend × Anzahl Imputationen) gemeldet.

Deskriptive Statistiken

Abbildung 5-12
Deskriptive Statistik für “tenure” (Beschäftigungsdauer)

Daten	Imputation	N	Mittelwert	Standardabweichung	Minimum	Maximum
Originaldaten		968	35,56	21,268	1,00	72,00
Imputierte Werte	1	32	36,06	24,218	-6,72	90,02
	2	32	37,64	22,229	-,19	88,03
	3	32	30,82	27,245	-40,99	104,77
	4	32	39,97	20,585	1,29	80,50
	5	32	37,87	20,669	3,44	94,21
Daten nach Imputation vervollständigen	1	1000	35,58	21,355	-6,72	90,02
	2	1000	35,63	21,291	-,19	88,03
	3	1000	35,41	21,484	-40,99	104,77
	4	1000	35,70	21,251	1,00	80,50
	5	1000	35,63	21,243	1,00	94,21

Die Tabellen “Deskriptive Statistik” zeigen Zusammenfassungen für Variablen mit imputierten Werten. Für jede Variable wird eine separate Tabelle erstellt. Die Typen der gezeigten Statistik hängen davon ab, ob die Variable metrisch oder kategorial ist.

Die Statistik für metrische Variablen umfasst Anzahl, Mittelwert, Standardabweichung, Minimum und Maximum, die für die Originaldaten, jedes Set an imputierten Werten und jedes vollständige Daten-Set (die Kombination aus Originaldaten und imputierten Werten) angezeigt werden.

Die Tabelle “Deskriptive Statistik” für *tenure (Beschäftigungsdauer)* zeigt Mittelwerte und Standardabweichungen in jedem Set von imputierten Werten, die ungefähr denen in den Originaldaten entsprechen. Es stellt sich jedoch ein unmittelbares Problem, wenn Sie sich das Minimum ansehen und sehen, dass die negativen Werte für *tenure* imputiert wurden.

Abbildung 5-13
Deskriptive Statistik für "marital" (Familienstand)

Daten	Im...	Ka...	N	Prozent
Originaldaten		0	456	51,5
		1	429	48,5
Imputierte Werte	1	0	51	44,3
		1	64	55,7
	2	0	41	35,7
		1	74	64,3
	3	0	49	42,6
		1	66	57,4
	4	0	43	37,4
		1	72	62,6
	5	0	53	46,1
		1	62	53,9
Daten nach Imputation vervollständigen	1	0	507	50,7
		1	493	49,3
	2	0	497	49,7
		1	503	50,3
	3	0	505	50,5
		1	495	49,5
	4	0	499	49,9
		1	501	50,1
	5	0	509	50,9
		1	491	49,1

Für kategoriale Variablen umfasst die Statistik Anzahl und Prozent nach Kategorie für die Originaldaten, imputierten Werte und vollständigen Daten. Die Tabelle für *marital (Familenstand)* hat ein interessantes Ergebnis, da für die imputierten Werte ein größerer Anteil der Fälle als in den Originaldaten als verheiratet geschätzt wurde. Hierbei könnte es sich um eine zufällige Variation handeln. Alternativ könnte die Möglichkeit des Fehlens in Zusammenhang mit dem Wert dieser Variable stehen.

Abbildung 5-14
Deskriptive Statistik für "income" (Haushaltseinkommen in Tausend)

Daten	Imputation	N	Mittelwert	Standardabweichung	Minimum	Maximum
Originaldaten		821	71,1462	83,14424	9,0000	944,0000
Imputierte Werte	1	179	87,6574	91,13179	-189,1959	373,2412
	2	179	101,6724	94,20599	-122,0010	346,4294
	3	179	100,9445	95,00789	-127,8572	342,5208
	4	179	107,0787	90,23638	-113,0959	369,9674
	5	179	101,1043	90,40865	-167,6978	314,2533
Daten nach Imputation vervollständigen	1	1000	74,1017	84,81851	-189,1959	944,0000
	2	1000	76,6104	85,98067	-122,0010	944,0000
	3	1000	76,4801	86,10024	-127,8572	944,0000
	4	1000	77,5781	85,52821	-113,0959	944,0000
	5	1000	76,5087	85,22154	-167,6978	944,0000

Wie *tenure* und alle anderen metrischen Variablen zeigt *income (Haushaltseinkommen in Tausend)* negative imputierte Werte — daher müssen wir ein angepasstes Modell mit Nebenbedingungen bei bestimmten Variablen einsetzen. *income* zeigt jedoch weitere mögliche Probleme. Die mittleren Werte für jede Imputation sind entscheidend höher als bei den Originaldaten und die

Maximumwerte für jede Imputation sind entscheidend niedriger als für die Originaldaten. Die Verteilung des Einkommens tendiert dazu, rechtslastig zu sein. Das könnte also die Ursache des Problems sein.

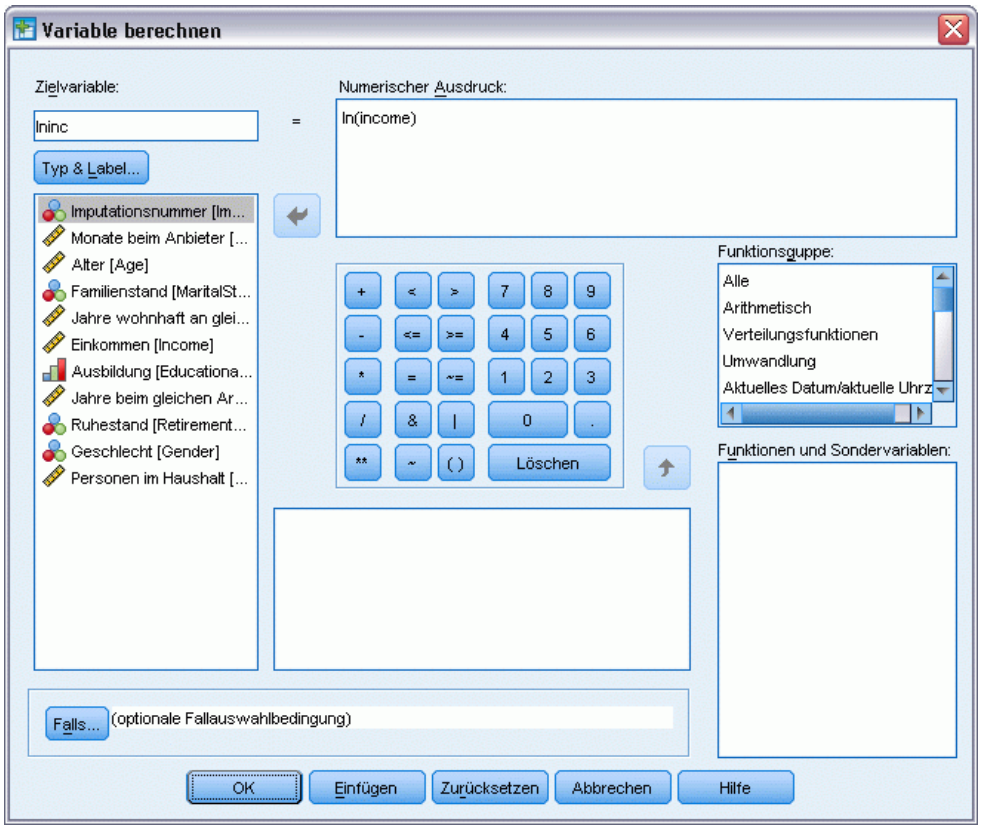
Angepasstes Imputationsmodell

Um zu verhindern, dass imputierte Werte außerhalb eines angemessenen Wertebereichs für jede Variable fallen, geben wir ein angepasstes Imputationsmodell mit Nebenbedingungen für die Variablen an. Zudem ist *Household income in thousands* (Haushaltseinkommen in Tausend) stark rechtslastig und die weitere Analyse wird wahrscheinlich den Logarithmus von *income* nutzen. Daher scheint die direkte Imputation von "log-income" Sinn zu ergeben.

- Stellen Sie sicher, dass das Original-Daten-Set aktiv ist.
- Wählen Sie zum Erstellen einer Variable "log-income" die folgenden Menübefehle aus:

Transformieren
  Variable berechnen...

Abbildung 5-15
Variable berechnen, Dialogfeld



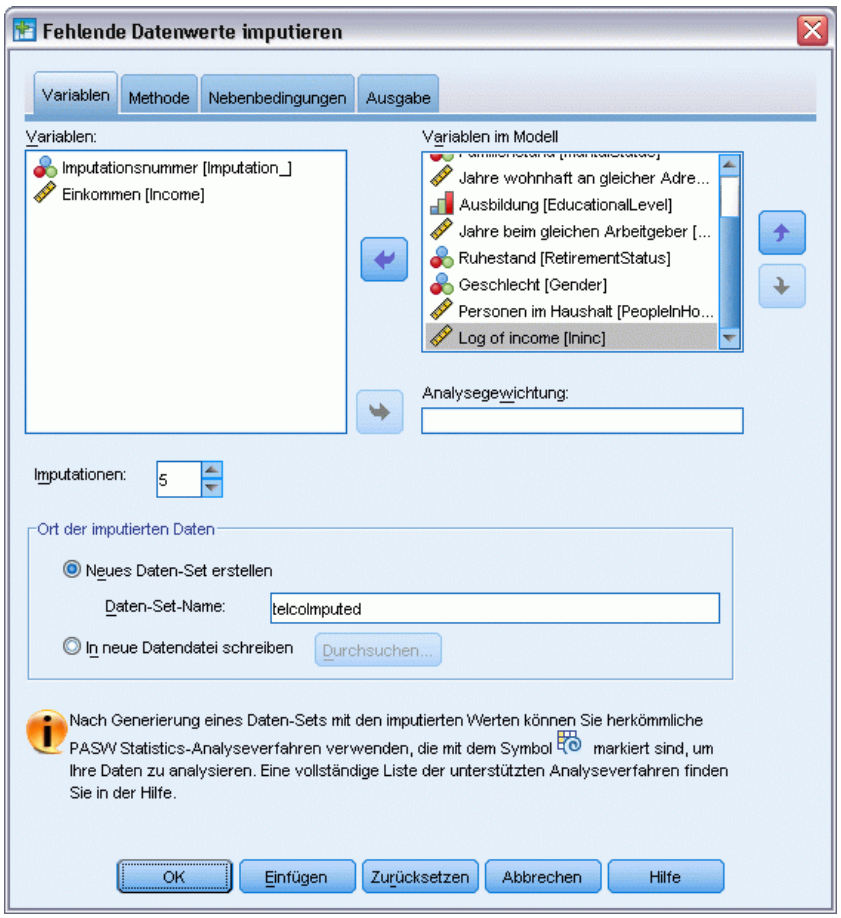
- Geben Sie *lninc* als Zielvariable ein.

- Geben Sie ln(income) als numerischen Ausdruck ein.
- Klicken Sie auf Typ & Label.

Abbildung 5-16
Typ und Label, Dialogfeld

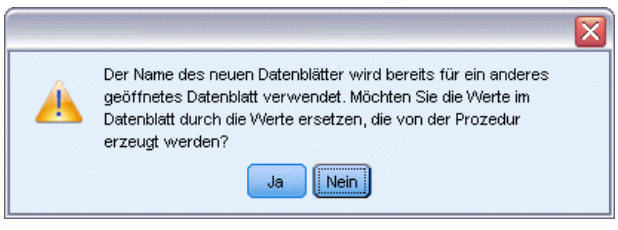
- Geben Sie *Log of income* als Label an.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Variable berechnen” auf OK.

Abbildung 5-17
Registerkarte "Variablen" mit "Log of income" als Ersatz für "Household income in thousands" im Imputationsmodell



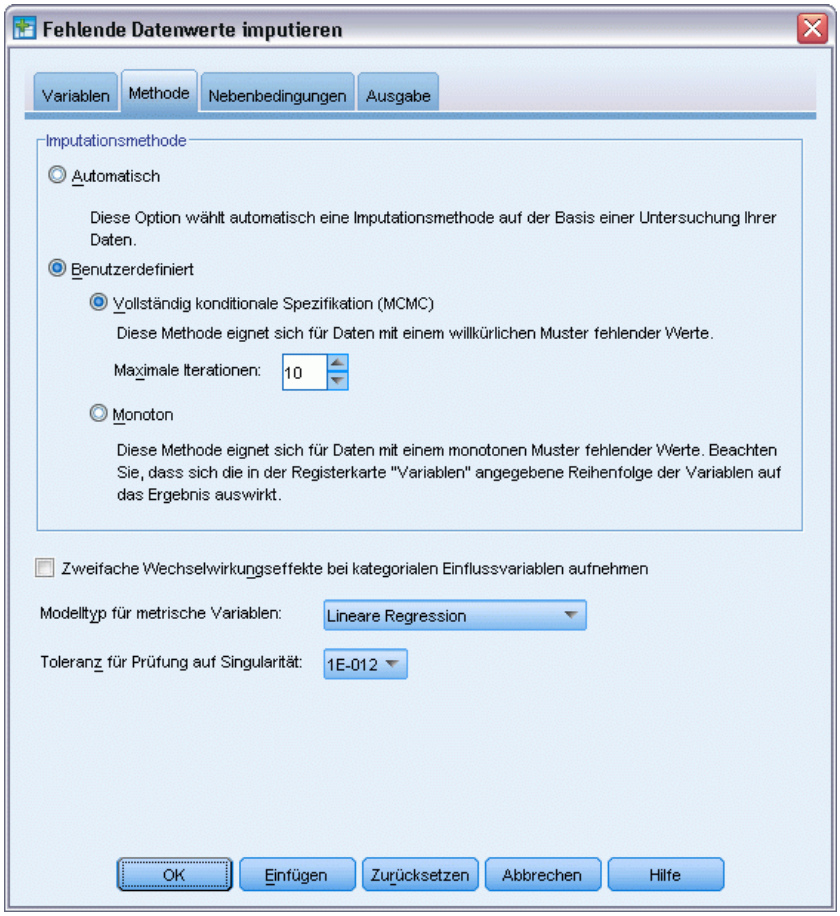
- Rufen Sie das Dialogfeld "Fehlende Datenwerte ersetzen" wieder auf und klicken Sie auf die Registerkarte Variablen.
- Deaktivieren Sie *Household income in thousands [income]* und wählen Sie *Log of income [lninc]* als Variablen im Modell.
- Klicken Sie auf die Registerkarte Methode.

Abbildung 5-18
Warnung über das Ersetzen eines bestehenden Daten-Sets


Der Name des neuen Datenblätter wird bereits für ein anderes geöffnetes Datenblatt verwendet. Möchten Sie die Werte im Datenblatt durch die Werte ersetzen, die von der Prozedur erzeugt werden?
Ja
Nein

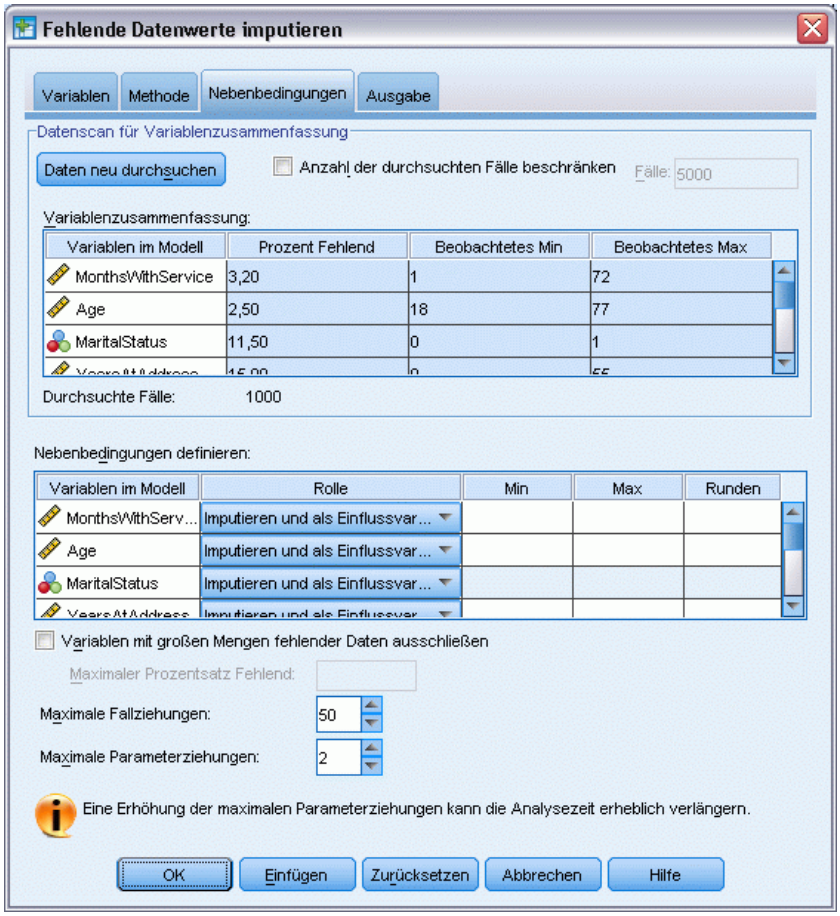
- Klicken Sie in der angezeigten Warnung auf Ja.

Abbildung 5-19
Registerkarte "Methode"



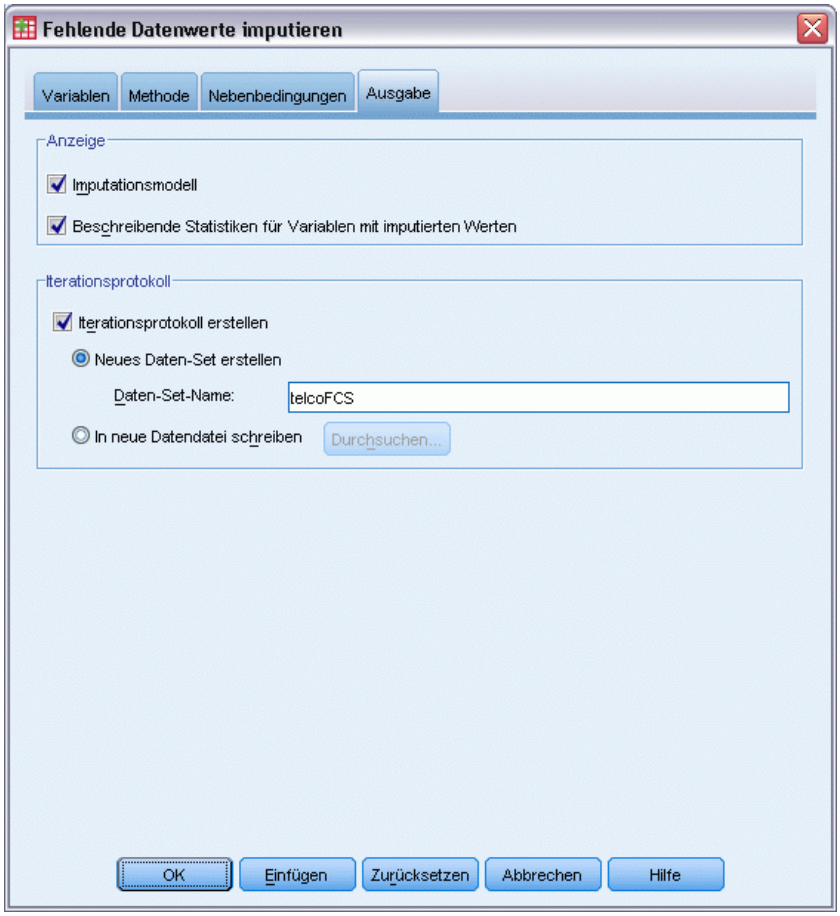
- Wählen Sie Benutzerdefiniert und belassen Sie Vollständig konditionale Spezifikation als Imputationsmethode aktiviert.
- Klicken Sie auf die Registerkarte Nebenbedingungen.

Abbildung 5-20
Nebenbedingungen, Registerkarte



- Klicken Sie auf Daten durchsuchen.
- Geben Sie im Raster “Nebenbedingungen definieren” 1 als Minimumwert für *Months with service [tenure]* ein.
- Geben Sie 18 als Minimumwert für *age (Age in years)* ein.
- Geben Sie 0 als Minimumwert für *address (Years at current address)* ein.
- Geben Sie 0 als Minimumwert für *employ (Years with current employer)* ein.
- Geben Sie 1 als Minimumwert und 1 als Rundungsstufe für *reside (Number of people in household)* ein. Beachten Sie, dass zwar viele der anderen metrischen Variablen als ganzzahlige Werte ausgewertet werden, es sich aber empfiehlt zu formulieren, dass jemand für 13,8 Jahre an seiner aktuellen Anschrift gewohnt hat, aber nicht wirklich anzunehmen, dass 2,2 Personen dort leben.
- Geben Sie 0 als Minimumwert für *lninc (Log of income)* ein.
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 5-21
Registerkarte "Ausgabe"



- Wählen Sie Iterationsprotokoll erstellen und geben Sie telcoFCS als Namen für das neue Daten-Set ein.
- Klicken Sie auf OK.

Imputationsnebenbedingungen

Abbildung 5-22
Imputationsnebenbedingungen

	Rolle in der Imputation		Imputierte Werte		
	Abhängig	Einflussvariable	Minimum	Maximum	Runden
Months with service	Ja	Ja	1	(ohne)	
Age in years	Ja	Ja	18	(ohne)	
Marital status	Ja	Ja			
Years at current address	Ja	Ja	0	(ohne)	
Level of education	Ja	Ja			
Years with current …	Ja	Ja	0	(ohne)	
Retired	Ja	Ja			
Gender	Ja	Ja			
Number of people in …	Ja	Ja	1	(ohne)	Ganzzahl
Lninc	Ja	Ja	0	(ohne)	

Das angepasste Imputationsmodell resultiert in einer neuen Tabelle, die die Nebenbedingungen für das Imputationsmodell zusammenfasst. Alles scheint Ihren Angaben zu entsprechen.

Deskriptive Statistik

Abbildung 5-23
Deskriptive Statistik für "tenure" (Beschäftigungsdauer)

Daten	Imputation	N	Mittelwert	Standardabweichung	Minimum	Maximum
Originaldaten		968	35,56	21,268	1,00	72,00
Imputierte Werte	1	32	37,90	17,621	6,40	86,94
	2	32	40,97	24,517	5,33	90,48
	3	32	37,57	19,913	9,97	93,52
	4	32	39,69	21,644	9,25	83,61
	5	32	39,85	20,093	7,21	81,37
Daten nach Imputation vervollständigen	1	1000	35,64	21,158	1,00	86,94
	2	1000	35,73	21,387	1,00	90,48
	3	1000	35,63	21,220	1,00	93,52
	4	1000	35,69	21,282	1,00	83,61
	5	1000	35,70	21,235	1,00	81,37

Die Tabelle "Deskriptive Statistik" für *tenure (Months with service)* für das angepasste Imputationsmodell mit Nebenbedingungen zeigt, dass das Problem negativer imputierter Werte für *tenure* gelöst wurde.

Abbildung 5-24
Deskriptive Statistik für "marital" (Familienstand)

Daten	Im...	Ka...	N	Prozent
Originaldaten		0	456	51,5
		1	429	48,5
Imputierte Werte	1	0	46	40,0
		1	69	60,0
	2	0	43	37,4
		1	72	62,6
	3	0	60	52,2
		1	55	47,8
	4	0	45	39,1
		1	70	60,9
	5	0	49	42,6
		1	66	57,4
Daten nach Imputation vervollständigen	1	0	502	50,2
		1	498	49,8
	2	0	499	49,9
		1	501	50,1
	3	0	516	51,6
		1	484	48,4
	4	0	501	50,1
		1	499	49,9
	5	0	505	50,5
		1	495	49,5

Die Tabelle für *marital (Marital status)* hat jetzt eine Imputation (3), deren Verteilung mehr den Originaldaten entspricht, die Mehrzahl zeigt aber im Vergleich zu den Originaldaten immer noch einen großen Anteil von Fällen, die als verheiratet geschätzt werden. Das könnte an der zufälligen Variation liegen, erfordert aber eventuell auch eine weitere Studie der Daten, um festzustellen, ob diese Werte nicht zufällig fehlen ("missing at random" - MAR). Dem gehen wir hier nicht weiter nach.

Abbildung 5-25
Deskriptive Statistik für lninc (Log of income)

Daten	Imputation	N	Mittelwert	Standardabweichung	Minimum	Maximum
Originaldaten		821	3,9291	,75305	2,1972	6,8501
Imputierte Werte	1	179	4,1816	,94574	1,4428	6,6748
	2	179	4,2562	,98346	1,6633	6,8224
	3	179	4,1743	1,01487	1,4443	6,8437
	4	179	4,1774	,82705	2,2532	6,2680
	5	179	4,1894	,96403	1,6667	6,6677
Daten nach Imputation vervollständigen	1	1000	3,9743	,79638	1,4428	6,8501
	2	1000	3,9876	,80842	1,6633	6,8501
	3	1000	3,9730	,81107	1,4443	6,8501
	4	1000	3,9735	,77228	2,1972	6,8501
	5	1000	3,9756	,80064	1,6667	6,8501

Wie *tenure* und alle anderen metrischen Variablen zeigt *lninc (Log of income)* keine negativen imputierten Werte. Ferner liegen die Mittelwerte für die Imputationen näher am Mittelwert für die Originaldaten als im automatischen Imputationslauf — für *income* beträgt der Mittelwert für die Originaldaten für *lninc* ungefähr $e^{3,9291}$=50,86, während der typische Mittelwert unter

den Imputationen ungefähr $e^{4,2}$=66,69 beträgt. Zusätzlich liegen die Maximumwerte für jede Imputation näher am Maximumwert für die Originaldaten.

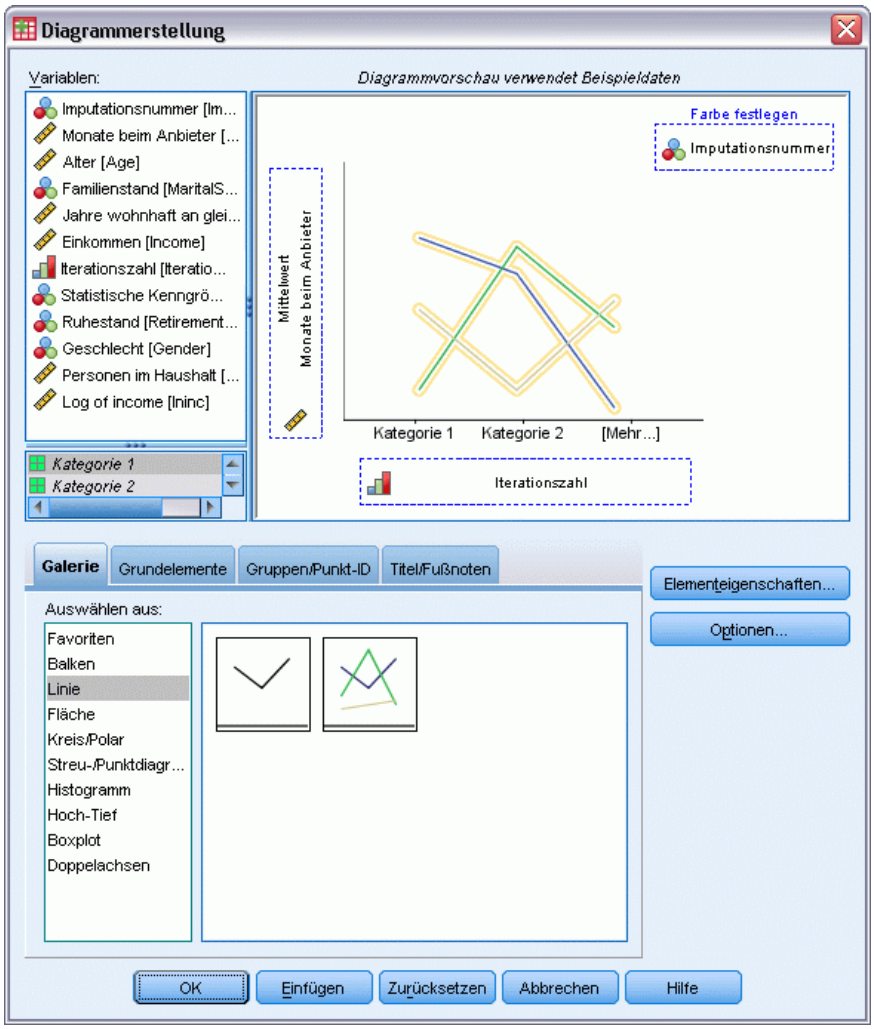
Prüfen auf FCS-Konvergenz

Wenn Sie die Methode der vollständig konditionalen Spezifikation verwenden, empfiehlt es sich, Darstellungen der Mittelwerte und Standardabweichungen je Iteration und Imputation für jede abhängige metrische Variable zu prüfen, für die Werte imputiert werden, um bei der Bewertung der Modellkonvergenz zu helfen.

- Um diese Art von Diagramm zu erstellen, aktivieren Sie das Daten-Set *telcoFCS* und wählen Sie dann aus den Menübefehlen:

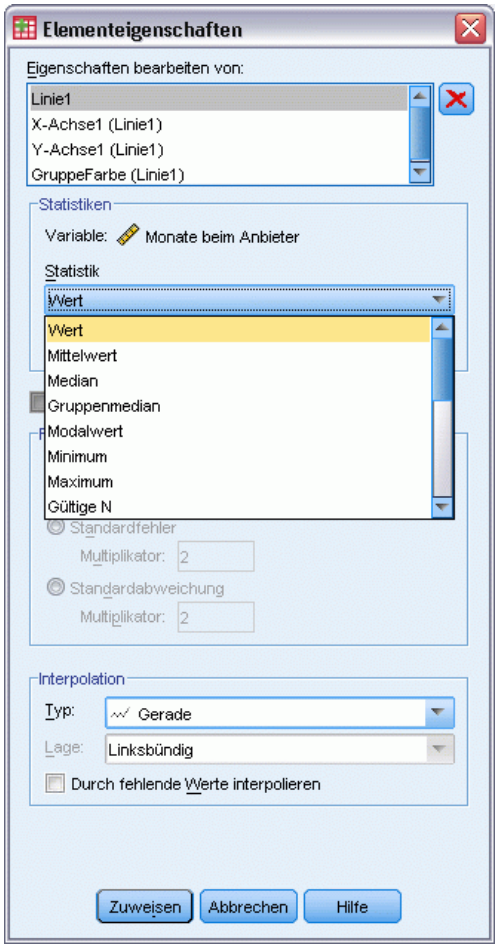
  Grafiken
  Diagrammerstellung...

Abbildung 5-26
Diagrammerstellung, Mehrere Linien, Diagramm



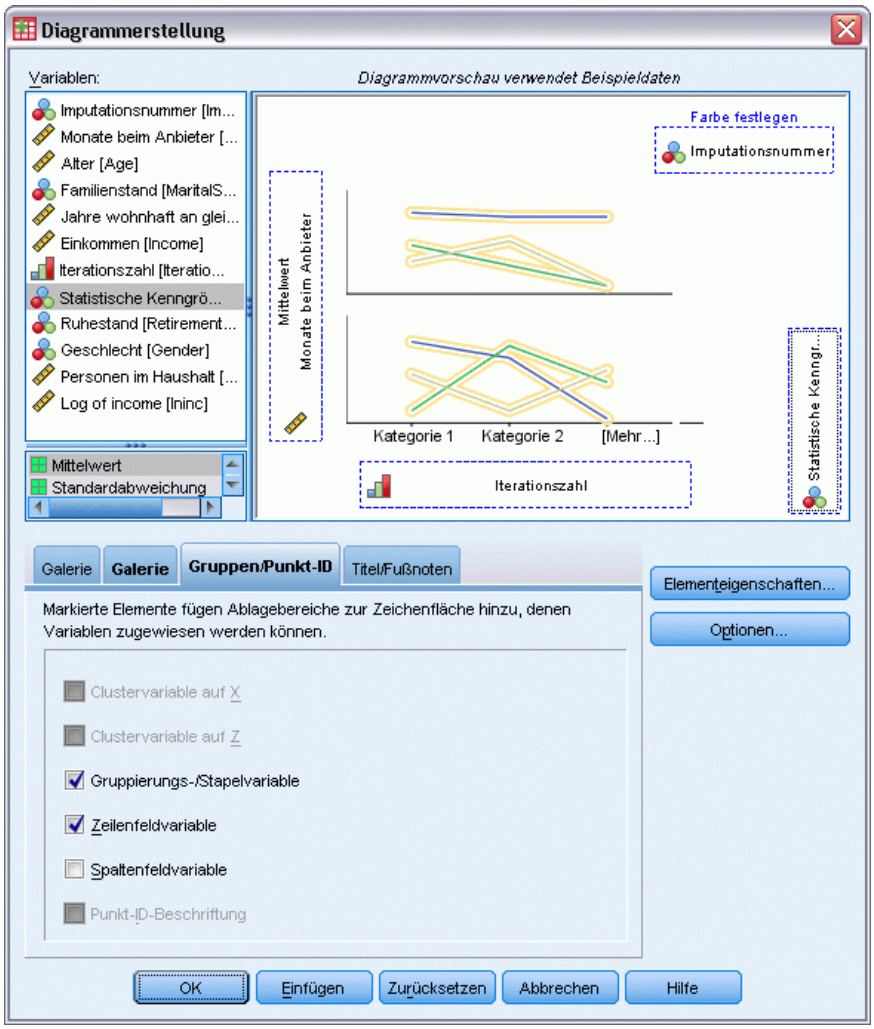
- Wählen Sie die Galerie Linien aus und wählen Sie “Mehrfachlinien”.
- Wählen Sie *Months with service [tenure]* als auf der *Y*-Achse darzustellende Variable.
- Wählen Sie *Iteration Number [Iteration_]* als auf der *X*-Achse darzustellende Variable aus.
- Wählen Sie *Imputationszahl [Imputationen_]* als Variable, um die Farben danach einzustellen.

Abbildung 5-27
Diagrammerstellung, Elementeigenschaften



- Wählen Sie in den Elementeigenschaften Wert als anzuzeigende Statistik.
- Klicken Sie auf Zuweisen.
- Klicken Sie in der Diagrammerstellung auf die Registerkarte Gruppen/Punkt-ID.

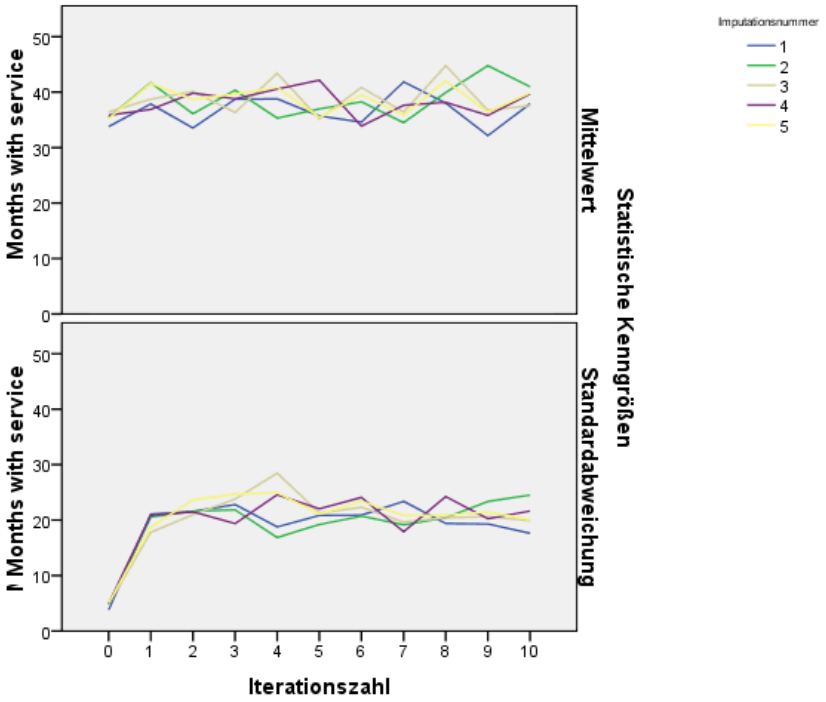
Abbildung 5-28
Diagrammerstellung, Registerkarte "Gruppen/Punkt-ID"



- Wählen Sie Zeilenfeldvariable.
- Wählen Sie *Auswertungsstatistik [SummaryStatistic_]* als Feldvariable.
- Klicken Sie auf OK.

FCS-Konvergenzdiagramme

Abbildung 5-29
FCS-Konvergenzdiagramm



Sie haben ein Paar von Mehrfachliniendiagrammen erstellt, die die mittlere und die Standardabweichung der imputierten Werte von *Months with service [tenure]* bei jeder Iiteration der FCS-Imputationmethode für jede der 5 angeforderten Imputationen anzeigen. Zweck dieser Darstellung ist, nach Mustern in den Linien zu suchen. Es sollte keine geben. Diese sehen geeignet “zufällig” aus. Sie können ähnliche Darstellungen für andere metrische Variablen erstellen. Beachten Sie, dass diese Darstellungen auch keine erkennbaren Muster zeigen.

Analyse vollständiger Daten

Jetzt scheinen Ihre imputierten Werte zufriedenstellend zu sein. Sie sind bereit, eine Analyse der “vollständigen” Daten durchzuführen. Das Daten-Set enthält eine Variable *Customer category [custcat]*, die den Kundenstamm nach Dienstnutzungsmustern segmentiert und die Kunden in vier Gruppen einteilt. Wenn Sie ein Modell mit demografischen Informationen anpassen können, um die Gruppenmitgliedschaft vorherzusagen, können Sie die Angebote für die einzelnen potenziellen Kunden anpassen.

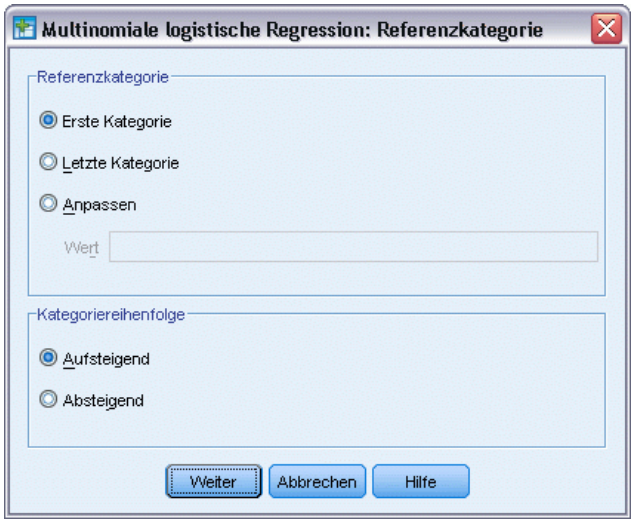
- Aktivieren Sie das Daten-Set *telcoImputed*. Um ein multinomiales logistisches Regressionsmodell für die vollständigen Daten zu erstellen, wählen Sie aus dem Menü:

  Analysieren
  Regression
  Multinomial logistisch...

Abbildung 5-30
Multinomiale logistische Regression, Dialogfeld

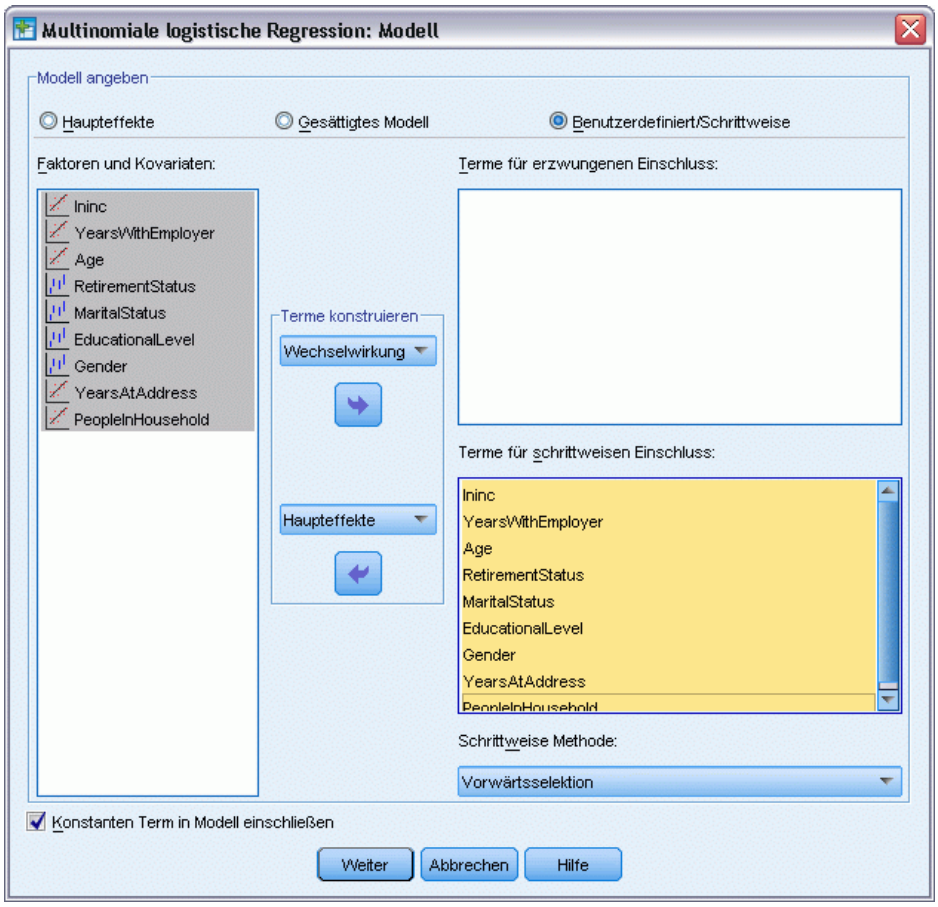
- Wählen Sie *Customer category* als abhängige Variable aus.
- Wählen Sie *Marital status*, *Level of education*, *Retired* und *Gender* als Faktoren.
- Wählen Sie *Age in Years*, *Years at current address*, *Years with current employer*, *Number of people in household* und *Log of income* als Kovariaten aus.
- Sie möchten andere Kunden mit denen vergleichen, die den Basisservice erhalten. Wählen Sie daher *Customer category* und klicken Sie auf Referenzkategorie.

Abbildung 5-31
Dialogfeld "Referenzkategorie"



- Wählen Sie Erste Kategorie.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multinomiale logistische Regression" auf Modell.

Abbildung 5-32
Dialogfeld "Modell"



- Wählen Sie Benutzerdefiniert/Schrittweise.
- Wählen Sie aus der Dropdown-Liste "Terme für schrittweisen Einschluss: Terme konstruieren" die Option Haupteffekte aus.
- Wählen Sie *Ininc* bis *reside* als schrittweise Terme aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multinomiale logistische Regression" auf OK.

Zusammenfassung der Schritte

Abbildung 5-33
Zusammenfassung der Schritte

Imputationsnummer	Modell	Aktion	Effekt(e)	Kriterien für die Modellanpassung	Effektauswahltests		
				-2 Log Likelihood	Chi-Quadrat[a]	Freiheitsgrade	Signifikanz
Originaldaten	0	Eingegeben	Intercept	1353,555	.		
	1	Eingegeben	ed	1260,972	92,583	12	,000
	2	Eingegeben	employ	1237,664	23,308	3	,000
	3	Eingegeben	marital	1229,808	7,856	3	,049
1	0	Eingegeben	Intercept	2762,531	.		
	1	Eingegeben	ed	2608,189	154,342	12	,000
	2	Eingegeben	employ	2563,671	44,518	3	,000
	3	Eingegeben	reside	2549,200	14,470	3	,002
	4	Eingegeben	address	2541,050	8,151	3	,043
2	0	Eingegeben	Intercept	2762,531	.		
	1	Eingegeben	ed	2603,940	158,591	12	,000
	2	Eingegeben	employ	2563,367	40,573	3	,000
	3	Eingegeben	marital	2545,743	17,624	3	,001
	4	Eingegeben	address	2536,532	9,211	3	,027
3	0	Eingegeben	Intercept	2762,531	.		
	1	Eingegeben	ed	2600,074	162,457	12	,000
	2	Eingegeben	employ	2558,560	41,514	3	,000
	3	Eingegeben	marital	2546,062	12,499	3	,006
	4	Eingegeben	address	2536,348	9,714	3	,021
4	0	Eingegeben	Intercept	2762,531	.		
	1	Eingegeben	ed	2601,616	160,915	12	,000
	2	Eingegeben	employ	2558,463	43,153	3	,000
	3	Eingegeben	marital	2543,747	14,716	3	,002
	4	Eingegeben	address	2533,341	10,406	3	,015
5	0	Eingegeben	Intercept	2762,531	.		
	1	Eingegeben	ed	2604,773	157,759	12	,000
	2	Eingegeben	employ	2561,792	42,980	3	,000
	3	Eingegeben	marital	2549,096	12,696	3	,005

Schrittweise Methode: Vorwärtsselektion

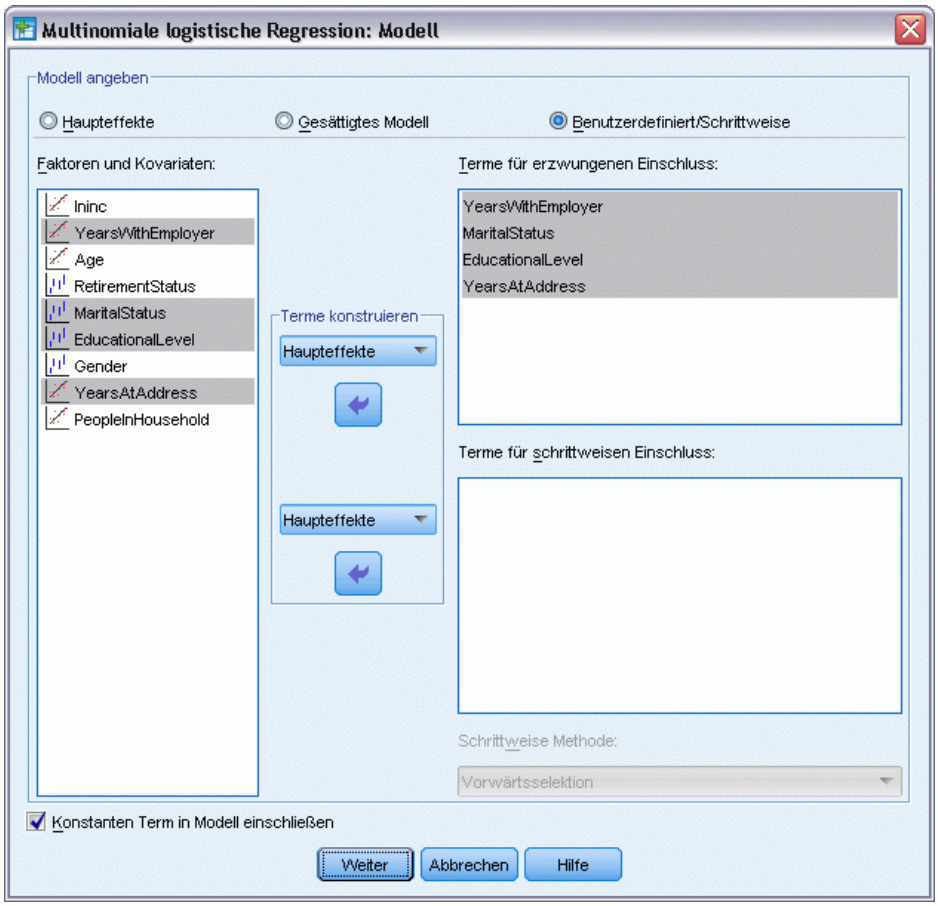
a. Das Chi-Quadrat für die Aufnahme beruht auf dem Likelihood-Quotienten-Test.

Die multinomiale logistische Regression unterstützt das Pooling von Regressionskoeffizienten. Sie werden jedoch feststellen, dass *alle* Tabellen in der Ausgabe die Ergebnisse für jede Imputation und die Originaldaten zeigen. Dies liegt an der Aufteilung der Datei bei *Imputation_*, so dass alle Tabellen, die die Aufteilungsvariable berücksichtigen, die Aufteilungsdateigruppen gemeinsam in einer einzigen Tabelle darstellen.

Sie werden ferner feststellen, dass die Tabelle “Parameterschätzer” keine gemeinsamen Schätzer zeigt. Sehen Sie sich hierzu die Zusammenfassung der Stufen an. Wir haben die schrittweise Auswahl von Modelleffekten angefordert und nicht für alle Imputationen wurde das gleiche Set an Effekten gewählt. Daher ist ein Pooling nicht möglich. Es werden dennoch hilfreiche Informationen bereitgestellt, da wir sehen, dass *ed (Level of education)*, *employ (Years with current employer)*, *marital (Marital status)* und *address (Years at current address)* regelmäßig durch die schrittweise Auswahl unter den Imputationen ausgewählt werden. Wir werden ein anderes Modell einsetzen, das genau diese Einflussvariablen verwendet.

Ausführen des Modells mit einer Untermenge an Einflussvariablen

Abbildung 5-34
Modell, Dialogfeld

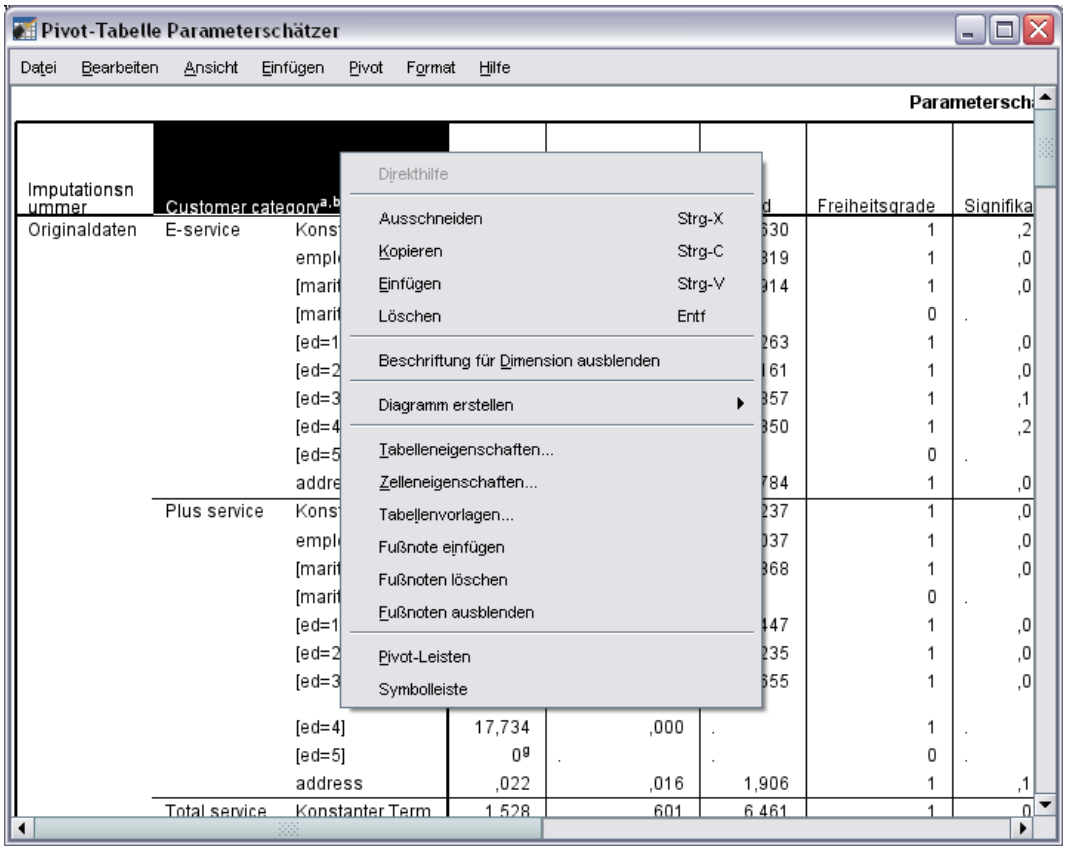


- Rufen Sie das Dialogfeld “Multinomiale logistische Regression” auf und klicken Sie auf Modell.
- Deaktivieren Sie die Variablen aus der Liste “Terme für schrittweisen Einschluss”.
- Wählen Sie aus der Dropdown-Liste “Terme für erzwungenen Einschluss: Terme konstruieren” die Option Haupteffekte aus.
- Wählen Sie *employ*, *marital*, *ed* und *address* als Terme für erzwungenen Einschluss.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Multinomiale logistische Regression” auf OK.

Gemeinsame Parameterschätzer

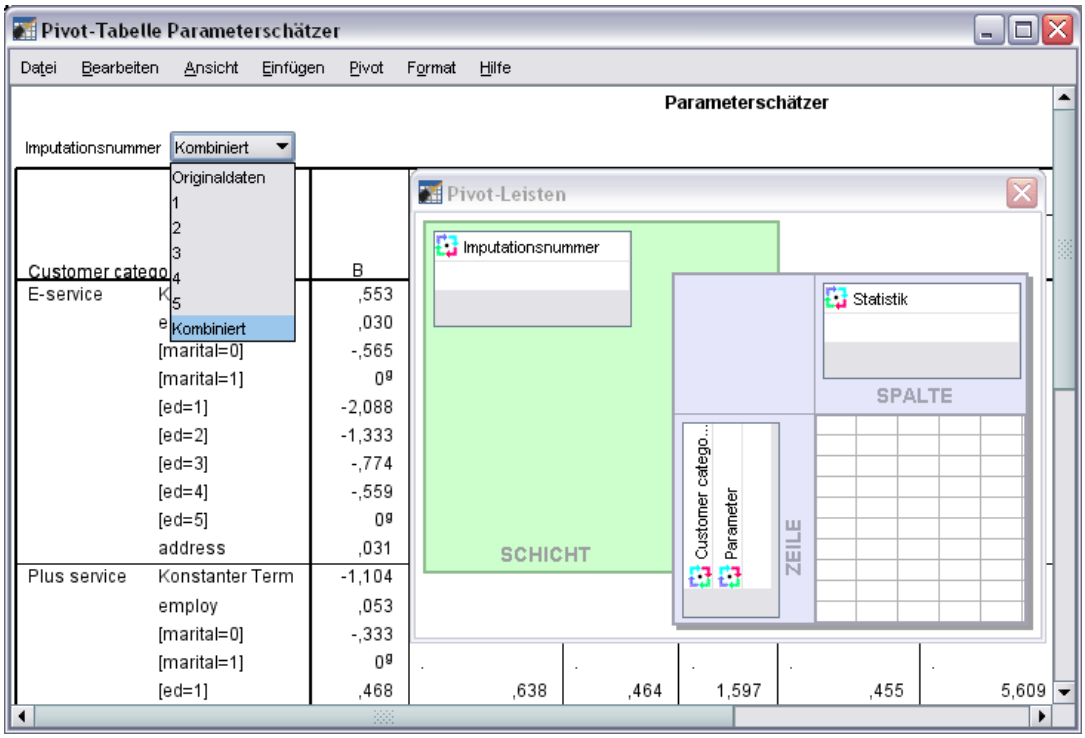
Diese Tabelle ist relativ groß, durch Pivotieren erhalten wir jedoch eine Reihe von unterschiedlichen, nützlichen Ansichten der Ausgabe.

Abbildung 5-35
Gemeinsame Parameterschätzer



- Aktivieren (doppelklicken) Sie die Tabelle und wählen Sie dann Pivot-Leisten aus dem Kontextmenü.

Abbildung 5-36
Gemeinsame Parameterschätzer



- Verschieben Sie die *Imputationsnummer* von der Zeile in die Schicht.
- Wählen Sie aus der Dropdown-Liste für “Imputationsnummer” Gemeinsam aus.

Abbildung 5-37
Gemeinsame Parameterschätzer

Imputationsnummer=Kombiniert

Customer category[a,b,c,d,e,f]		B	Standard fehler	Signifikanz	Exp(B)	95% Konfidenzintervall für Exp (B)		Anteil fehlende Info.	Relative Zunahme varianz	Relative Effizienz
						Untergrenze	Obergrenze			
E-service	Konstanter Term	,553	,435	,204				,076	,080	,985
	employ	,030	,012	,014	1,030	1,006	1,054	,051	,052	,990
	[marital=0]	-,565	,198	,004	,568	,385	,839	,076	,080	,985
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-2,088	,479	,000	,124	,048	,317	,079	,082	,985
	[ed=2]	-1,333	,454	,004	,264	,108	,644	,092	,098	,982
	[ed=3]	-,774	,458	,092	,461	,187	1,134	,075	,079	,985
	[ed=4]	-,559	,466	,231	,572	,229	1,428	,109	,116	,979
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,031	,012	,009	1,032	1,008	1,056	,140	,153	,973
Plus service	Konstanter Term	-1,104	,632	,082				,139	,152	,973
	employ	,053	,011	,000	1,054	1,032	1,076	,060	,062	,988
	[marital=0]	-,333	,179	,063	,717	,505	1,018	,011	,012	,998
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	,468	,638	,464	1,597	,455	5,609	,126	,136	,975
	[ed=2]	,702	,637	,272	2,017	,575	7,077	,140	,153	,973
	[ed=3]	,660	,644	,306	1,936	,546	6,867	,118	,127	,977
	[ed=4]	,452	,667	,499	1,571	,421	5,862	,160	,178	,969
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,015	,011	,157	1,016	,994	1,038	,118	,127	,977
Total service	Konstanter Term	1,086	,412	,009				,058	,060	,989
	employ	,042	,012	,001	1,043	1,018	1,068	,069	,071	,986
	[marital=0]	-,659	,201	,001	,517	,349	,767	,090	,095	,982
	[marital=1]	0[g]	.	.	.	.	.	.	.	.
	[ed=1]	-3,492	,534	,000	,030	,011	,087	,098	,104	,981
	[ed=2]	-1,772	,433	,000	,170	,073	,398	,075	,078	,985
	[ed=3]	-1,307	,441	,003	,271	,114	,643	,064	,066	,987
	[ed=4]	-,488	,432	,259	,614	,263	1,432	,076	,079	,985
	[ed=5]	0[g]	.	.	.	.	.	.	.	.
	address	,013	,013	,320	1,013	,987	1,039	,211	,244	,960

Diese Ansicht zeigt alle Statistikwerte für die gemeinsamen Ergebnisse. Sie können diese Koeffizienten auf die gleiche Art verwenden und interpretieren, wie Sie diese Tabelle für ein Daten-Set ohne fehlende Werte verwenden würden.

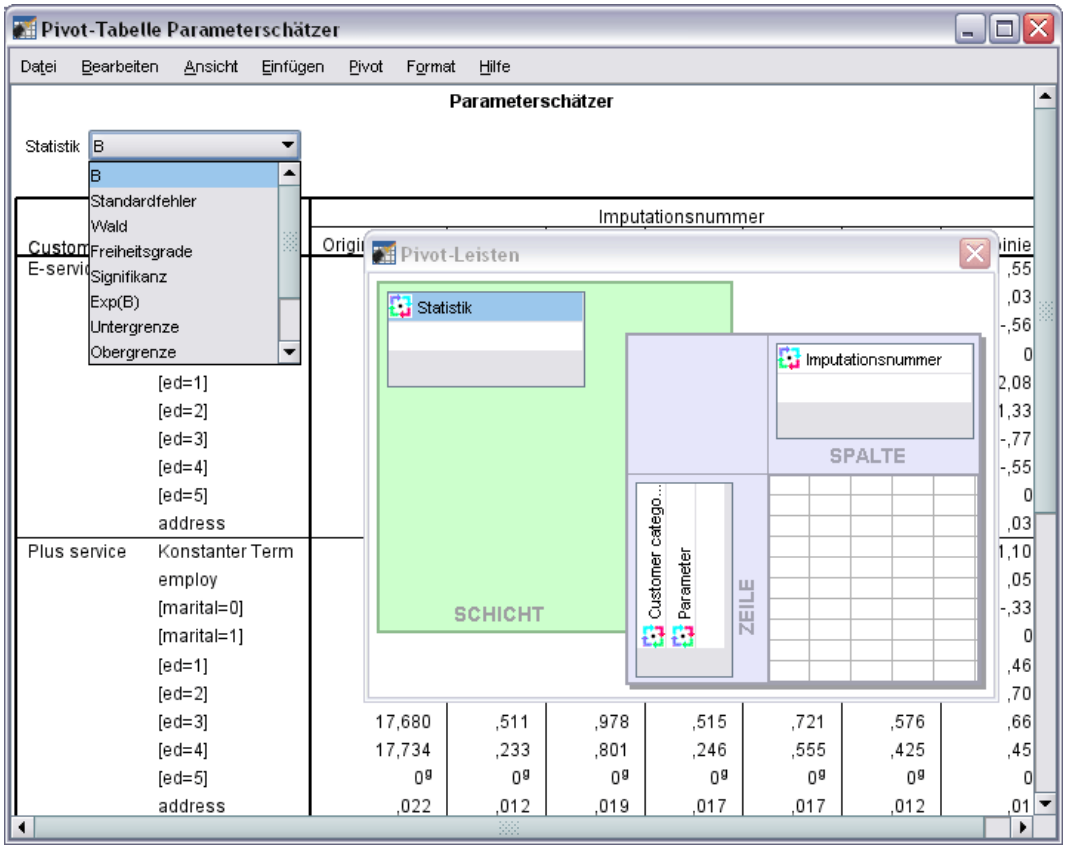
Die Tabelle der Parameterschätzer fasst den Effekt der einzelnen Einflussvariablen zusammen. Der Quotient des Koeffizienten zu seinem Standardfehler ergibt quadriert die Wald-Statistik. Wenn das Signifikanzniveau der Wald-Statistik gering ausfällt (kleiner als 0,05), ist der Parameter von 0 verschieden.

- Parameter mit signifikanten negativen Koeffizienten verringern die Likelihood dieser Antwortkategorie in Bezug auf die Referenzkategorie.
- Parameter mit positiven Koeffizienten erhöhen die Likelihood dieser Antwortkategorie.
- Die mit der letzten Kategorie jedes Faktors verbundenen Parameter sind mit konstantem Term redundant.

Es gibt drei zusätzliche Spalten in der Tabelle, die weitere Informationen für die gemeinsame Ausgabe bereitstellen. **Bruchteil der fehlenden Informationen** ist eine Schätzung des Verhältnisses fehlender Informationen zu “vollständigen” Informationen, basierend auf dem **relativen Anstieg der Varianz** aufgrund von Nichtantworten, das wiederum ein (modifiziertes) Verhältnis der Zwischenimputation und der durchschnittlichen Innenimputationsvarianz des

Regressionskoeffizienten ist. Die **relative Effizienz** ist ein Vergleich dieser Schätzung mit einer (theoretischen) Schätzung, die mit einer infiniten Anzahl von Imputationen berechnet wurde. Die relative Effizienz wird durch den Bruchteil der fehlenden Informationen und der Anzahl der Imputationen berechnet, die für das gemeinsame Ergebnis verwendet wurden. Wenn der Bruchteil der fehlenden Informationen groß ist, ist eine größere Anzahl von Imputationen erforderlich, um die relative Effizienz näher an 1 und die gemeinsame Schätzung näher an die idealisierte Schätzung zu bringen.

Abbildung 5-38
Gemeinsame Parameterschätzer



- Aktivieren (doppelklicken) Sie jetzt wieder die Tabelle und wählen Sie dann Pivot-Leisten aus dem Kontextmenü.
- Verschieben Sie die *Imputationsnummer* von der Schicht in die Spalte.
- Verschieben Sie *Statistik* von der Spalte in die Schicht.
- Wählen Sie aus der Dropdown-Liste “Statistik” B aus.

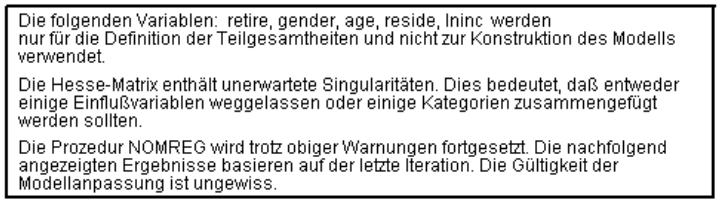
Abbildung 5-39
Gemeinsame Parameterschätzer, Imputationsnummer in Spalten und Statistik in Schicht

Statistik= B

Customer category[a,b,c,d,e,f]		Imputationsnummer						
		Originaldaten	1	2	3	4	5	Kombiniert
E-service	Konstanter Term	,817	,605	,366	,613	,627	,555	,553
	employ	,051	,033	,030	,028	,027	,030	,030
	[marital=0]	-,760	-,497	-,621	-,553	-,604	-,551	-,565
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-2,376	-2,171	-1,928	-2,111	-2,223	-2,007	-2,088
	[ed=2]	-1,754	-1,435	-1,131	-1,412	-1,383	-1,302	-1,333
	[ed=3]	-,943	-,870	-,584	-,816	-,837	-,764	-,774
	[ed=4]	-,798	-,610	-,353	-,665	-,678	-,491	-,559
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,030	,028	,033	,032	,036	,027	,031
Plus service	Konstanter Term	-17,768	-,915	-1,425	-,947	-1,191	-1,041	-1,104
	employ	,054	,055	,049	,051	,052	,054	,053
	[marital=0]	-,542	-,332	-,352	-,343	-,330	-,306	-,333
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	17,479	,287	,792	,317	,503	,442	,468
	[ed=2]	17,262	,503	1,022	,553	,800	,630	,702
	[ed=3]	17,680	,511	,978	,515	,721	,576	,660
	[ed=4]	17,734	,233	,801	,246	,555	,425	,452
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,022	,012	,019	,017	,017	,012	,015
Total service	Konstanter Term	1,528	1,157	,942	1,153	1,118	1,058	1,086
	employ	,038	,041	,039	,046	,040	,044	,042
	[marital=0]	-,459	-,647	-,749	-,612	-,666	-,624	-,659
	[marital=1]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	[ed=1]	-3,664	-3,406	-3,390	-3,727	-3,555	-3,380	-3,492
	[ed=2]	-2,305	-1,796	-1,594	-1,818	-1,878	-1,776	-1,772
	[ed=3]	-1,295	-1,347	-1,146	-1,398	-1,366	-1,276	-1,307
	[ed=4]	-,710	-,548	-,335	-,578	-,562	-,417	-,488
	[ed=5]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]	0[g]
	address	,005	,011	,018	,008	,019	,008	,013

Diese Ansicht der Tabelle empfiehlt sich für den Vergleich von Werten zwischen Imputationen, um eine schnelle optische Prüfung der Variation im Regressionskoeffizienten von Imputation zu Imputation und auch gegenüber den Originaldaten durchzuführen. Speziell durch das Umschalten der Statistik in der Schicht auf Standardfehler können Sie sehen, wie multiple Imputation die Variabilität in den Koeffizientenschätzungen im Vergleich zum listenweisen Ausschluss (Originaldaten) verringert hat.

Abbildung 5-40
Warnungen


Die folgenden Variablen: retire, gender, age, reside, lninc werden nur für die Definition der Teilgesamtheiten und nicht zur Konstruktion des Modells verwendet.
Die Hesse-Matrix enthält unerwartete Singularitäten. Dies bedeutet, daß entweder einige Einflußvariablen weggelassen oder einige Kategorien zusammengefügt werden sollten.
Die Prozedur NOMREG wird trotz obiger Warnungen fortgesetzt. Die nachfolgend angezeigten Ergebnisse basieren auf der letzte Iteration. Die Gültigkeit der Modellanpassung ist ungewiss.

In diesem Beispiel verursacht das Original-Daten-Set jedoch einen Fehler, der die großen Parameterschätzer für den konstanten Term *Plus service* und die nicht redundanten Stufen von *ed (Level of education)* in der Originaldatenspalte der Tabelle erklärt.

Auswertung

Unter Verwendung der Verfahren multipler Imputation haben Sie Muster fehlender Werte analysiert und festgestellt, dass viele Informationen vermutlich verloren gehen würden, wenn ein einfach listenweiser Ausschluss verwendet werden würde. Nach einem ersten automatischen Durchlauf der multiplen Imputation haben Sie festgestellt, dass Nebenbedingungen benötigt werden, um imputierte Werte in einem vernünftigen Rahmen zu halten. Der Durchlauf mit Nebenbedingungen sorgte für gute Ergebnisse und es gab keinen direkten Nachweis, dass die FCS-Methode nicht konvergiert hat. Unter Verwendung des “vollständigen” Daten-Sets mit mehrfach imputierten Werten haben Sie eine multinomiale logistische Regression an die Daten angepasst und gemeinsame Regressionsschätzer erhalten. Zudem haben Sie erkannt, dass die abschließende Modellanpassung tatsächlich mittels listenweisen Ausschlusses an den Originaldaten nicht möglich gewesen wäre.

Anhang

A

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses. Für jeder der folgenden Sprachen gibt es einen eigenen Ordner innerhalb des Unterverzeichnisses “Samples”: Englisch, Französisch, Deutsch, Italienisch, Japanisch, Koreanisch, Polnisch, Russisch, Vereinfachtes Chinesisch, Spanisch und Traditionelles Chinesisch.

Nicht alle Beispieldateien stehen in allen Sprachen zur Verfügung. Wenn eine Beispieldatei nicht in einer Sprache zur Verfügung steht, enthält der jeweilige Sprachordner eine englische Version der Beispieldatei.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien.

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.

- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.
- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel () wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = “ausgesprochen angemessen” bis 9 = “ausgesprochen unangemessen” zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie () wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = “am meisten bevorzugt” bis 15 = “am wenigsten bevorzugt” zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von “Overall preference” (Allgemein bevorzugt) bis “Snack, with beverage only” (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, “Overall preference” (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.

- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.
- **car_insurance_claims.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **car_sales_uprepared.sav.** Hierbei handelt es sich um eine modifizierte Version der Datei *car_sales.sav*, die keinerlei transformierte Versionen der Felder enthält.
- **carpet.sav.** In einem beliebten Beispiel möchte einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung setzt sich aus drei Faktorenebenen zusammen, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Ebenen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur “Saisonale Zerlegung” berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.

- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.
- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als “AA”, “BB”, “CC”, “DD”, “EE” und “FF” bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customer_information.sav.** Eine hypothetische Datendatei mit Kundenmailingdaten wie Name und Adresse.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.
- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.

- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.
- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dmdata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen über Demografie und Einkäufe für ein Direktmarketing-Unternehmen enthält.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der “Stillman-Diät” . Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.
- **german_credit.sav.** Diese Daten sind aus dem Daten-Set “German credit” im Repository of Machine Learning Databases () an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.

- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell () legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman () verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: “crowds” (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), “audience” (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), “public” (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), “mobs” (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), “primary groups” (Primärgruppen, vertraulich), “secondary groups” (Sekundärgruppen, freiwillig) und “modern community” (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).
- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insurance_claims.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen möchte. Jeder Fall entspricht einem Anspruch.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.

- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs "Quellen" erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.
- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener*(Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen (,) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.

- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.

- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem frühren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Eine hypothetische Datendatei mit Kauftransaktionsdaten wie Kaufdatum, gekauften Artikeln und Geldbetrag für jede Transaktion.
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln ().
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle

der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.

- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als "good" (gut) "fair" (mittelmäßig) oder "poor" (schlecht) ab.
- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre () vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option "Data Preparation" bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.

- **survey_sample.sav.** Diese hypothetische Datendatei enthält Umfragedaten einschließlich demografischer Daten und verschiedener Meinungskennzahlen.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.
- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei ist eine Untermenge der Datendatei *telco.sav*, allerdings wurde ein Teil der demografischen Datenwerte durch fehlende Werte ersetzt.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das

Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.

- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle () vorgestellt und analysiert.
- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle () vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage (). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder (). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Index

Analyse fehlender Werte, 3, 37
  Deskriptive Statistik, 37
  Deskriptive Statistiken, 8
  EM, 10
  Erwartungs-Maximierung, 13
  Imputieren fehlender Werte, 9
  MCAR-Test, 10
  Methoden, 9
  Muster, 6, 45
  Regression, 12
  Schätzen von Statistiken, 9
  zusätzliche Funktionen beim Befehl, 14

Beispieldateien
  Lage, 85

EM
  in “Analyse fehlender Werte”, 10

Fälle in Tabellen
  in “Analyse fehlender Werte”, 6
FCS-Konvergenzdiagramm
  bei multipler Imputation, 73
Fehlende Datenwerte imputieren, 18
  Ausgabe, 24
  Imputationsmethode, 20
  Nebenbedingungen, 22
Fehlende Werte
  Univariate Statistiken, 8, 40
fehlende Werte, Muster, 47

gemeinsame Ergebnisse
  bei multipler Imputation, 73
gemeinsame Schätzer
  bei multipler Imputation, 79

Häufigkeiten extremer Werte
  in “Analyse fehlender Werte”, 8
Häufigkeitstabellen
  in “Analyse fehlender Werte”, 8

Indikatorvariablen
  in “Analyse fehlender Werte”, 8
Indikatorvariablen für fehlende Werte
  in “Analyse fehlender Werte”, 8
Iterationsprotokoll
  in Multiple Imputation, 24

Korrelationen
  in “Analyse fehlender Werte”, 10, 12
Kovarianz
  in “Analyse fehlender Werte”, 10, 12

Listenweiser Ausschluß
  in “Analyse fehlender Werte”, 3

MCAR-Test
  in “Analyse fehlender Werte”, 3
  in “Analyse fehlender Werte”, 48
MCAR-Test nach Little, 10
  in “Analyse fehlender Werte”, 3
  in “Analyse fehlender Werte”, 48
Mittelwert
  in “Analyse fehlender Werte”, 8, 10, 12
monotone Imputation
  in Multiple Imputation, 20
Multiple Imputation, 15, 25, 29, 50
  Deskriptive Statistik, 59, 67
  FCS-Konvergenzdiagramm, 73
  Fehlende Datenwerte ersetzen, 18
  fehlende Werte, Muster, 53
  gemeinsame Ergebnisse, 73
  gemeinsame Schätzer, 79
  Gesamtzusammenfassung der fehlenden Werte, 51
  Imputationsergebnisse, 58
  Imputationsspezifikationen, 57
  Modelle, 58
  Muster analysieren, 17
  Nebenbedingungen, 67
  Optionen, 34
  Variablenauswertung, 52
Muster analysieren, 17

Nichtübereinstimmung
  in “Analyse fehlender Werte”, 8
Normale Variaten
  in “Analyse fehlender Werte”, 12

Optionen
  Multiple Imputation, 34

Paarweiser Ausschluss
  in “Analyse fehlender Werte”, 3

Regression
in “Analyse fehlender Werte”, 12
Residuen
in “Analyse fehlender Werte”, 12

Sortieren von Fällen
in “Analyse fehlender Werte”, 6
Standardabweichung
in “Analyse fehlender Werte”, 8
Student-T-Test
in “Analyse fehlender Werte”, 12
in “Analyse fehlender Werte”, 41

t-Test
in “Analyse fehlender Werte”, 8
T-Test
in “Analyse fehlender Werte”, 41
Tabellarische Darstellung von Kategorien
in “Analyse fehlender Werte”, 8
in “Analyse fehlender Werte”, 42

Univariate Statistiken
in “Analyse fehlender Werte”, 40
Unvollständige Daten
siehe Analyse fehlender Werte, 3

vollständig konditionale Spezifikation
in Multiple Imputation, 20